**Project Objective:** to perform data analysis on the annual upstream flaring volumes as reported by satellite data over *Saudi Arabia.*

**Data type:** the data is a collected on an annual basis and consist of the following variables:

* **Country Name**
* **ISO Code**: country identifier code
* **Flare ID**: not a unique ID – and changes each year
* **Latitude:** of the detected flare
* **Longitude**: of the detected flare
* **Detection Frequency**: as a % out of 100
* **BCM**: flared gas volume
* **Avg. Temperature:** of the burning flare
* **Ellipticity:** a measure of flare shape
* **Type:** identifier of the source of flare

****

**Working dataset:**

* **Upload** only “flare upstream” sheet *for each year* from the attached excel files.
* **Filter** the data by country code (“SAU” & “SAUKWTNZ”)
* **Select** only (“ISO\_Code”, “Latitude”, “Longitude”, “Detection\_Frequency”, “BCM”)
  + **Rename columns to the above standard column names if different**

**Mutate** ID = seq.int(nrow()),

* Year\_FD = year of detection (i.e 2017, 2018, 2019, 2020 or 2021)

MMscf = ,

MMscf\_Day = ,

Days\_On\_Yr = Detection\_Frequency% ,

Days\_On\_Mnth = Days\_On\_Yr )

* **Arrange columns so it appears in the following order for *each year* ::**

ISO\_Code, ID, Year, Latitude, Longitude, Detection\_Frequency, BCM, MMscf, MMscf\_Day, Days\_On, Days\_On\_Mnth

*The working data set by this stage* is **5 tables** for the years from 2017 to 2021 and each **with 11 columns** as listed above. The number of rows varies for each year, and this is the challenge in the data.

**Challenges in *the Data*:**

1. The flare identifier ID (for each row) changes from year to year and thus can’t be used to identify repeat locations.
2. The satellite Latitude and Longitude coordinates are not perfect either and changes slightly even for a single location from one year to another.

**Project tasks and deliverables:**

**Task-1: find a way to identify repeat locations, set a unique ID for each location, and aggregate yearly data for each location.**

**Approach:**

* **First**: starting with Year-1 dataset, check the distance of each of Year-2 coordinates to Year-1 and use the results to generate key columns to help future queries. Example:
  + Distance between two GPS coordinates can be calculated using specialized geo-distance package or using *nested loops* to identify:
    - **Minimum distance**:
    - **Corresponding ID** 
      * help you identify the matching flare volumes in Year-2 to the GPS location in Year-1, and so on.
  + Minimum distance between the two GPS locations will help identify Repeat\_Status: Case\_when(
    - **0 km ~ Repeat\_Status = “Origin” or “O”**
    - **(0 to 1] km ~ Repeat\_Status = “Confirmed” or “C”**
    - **(1 to 1.6) km ~ Repeat\_Status = “Possible” or “P”**
    - **[1.6 to 2) km ~ Repeat\_Status = “Suspect” or “S”**
    - **>= 2 km ~ Repeat\_Status = “Unique” or “U”**
* **Second:** the columns of only Repeat\_Status = c(“O”, “C” , “P” , “S”) are ***left\_joined*** together in a Meta\_table under one unique-ID, Latitude and Longitude with year of first detection(Year\_FD) shown. There should be only one row for each GPS location where the values in Year-1, 2, 3, 4 and 5 correspond to the original table columns (**Detection\_Frequency**, **BCM**, **MMscf**, **MMscf\_Day**, **Days\_On\_Yr**, **Days\_On\_Mnth**, and **Repeat\_status**)\*

**\***If needed add an identifier for each year data. For example BCM\_2017, BCM\_2018 .. etc

**\*** Any rows with no corresponding values in the year columns should display 0 zero values.

Ex: unique location only detected in year-3 and have no “C”, “P” or “S” status to previous locations should show zero values for the columns in year-1 and year-2. Similarly, any location detected in year1 and 2 but no confirmed detection in future years should display zero values**.**

* **Third**: the output should be a Meta\_table which display unique rows of each location, with unique ID. The columns include:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unique-ID | ISO\_Code | Year\_FD | Latitude | Longitude | Year-1 | Year-2 | Year-3 | Year-4 | Year-5 |  |
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |
| N |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

**EXAMPLE:**

1. **Location detected in SAU first in year-1, and confirmed detection in year-2, year-3, possibly year-4 and suspected in year-5**
2. **Location detected in SAUKWNZ first in year 2, and confirmed detection in year 3 and 4, possible repeat in year 5**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unique-ID** | **ISO\_Code** | **Year\_FD** | **Latitude** | **Longitude** | **Detection\_Frequency\_2017** | **BCM\_2017** | **MMscf\_2017** | **MMscf\_Day\_2017** | **Days\_On\_Yr\_2017** | **Days\_On\_Mnth\_2017** | **Repeat\_Status\_2017** | **……** | **Repeat\_Status\_2018** | **……** | **Repeat\_Status\_2019** |  | **Repeat\_Status\_2019** |  | **Repeat\_Status\_2019** |
| **1** | **SAU** | **2017** | **25.65321** | **45.34521** | 95.2 | 1.2 | 1200 | 3.287 | 374.7 | 28.97 | Unique or “U” or Origin “O” |  | Confirmed or “C” |  | Confirmed or “C” |  | Possible or “P” |  | Suspect or “S” |
| **2** | **SAUKWNZ** | **2017** | **26.65321** | **44.34521** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  | Unique or “U” or Origin “O” |  | Confirmed or “C” |  | Confirmed or “C” |  | Possible or “P” |
| 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**End.**

Thank you,

Fahad BH