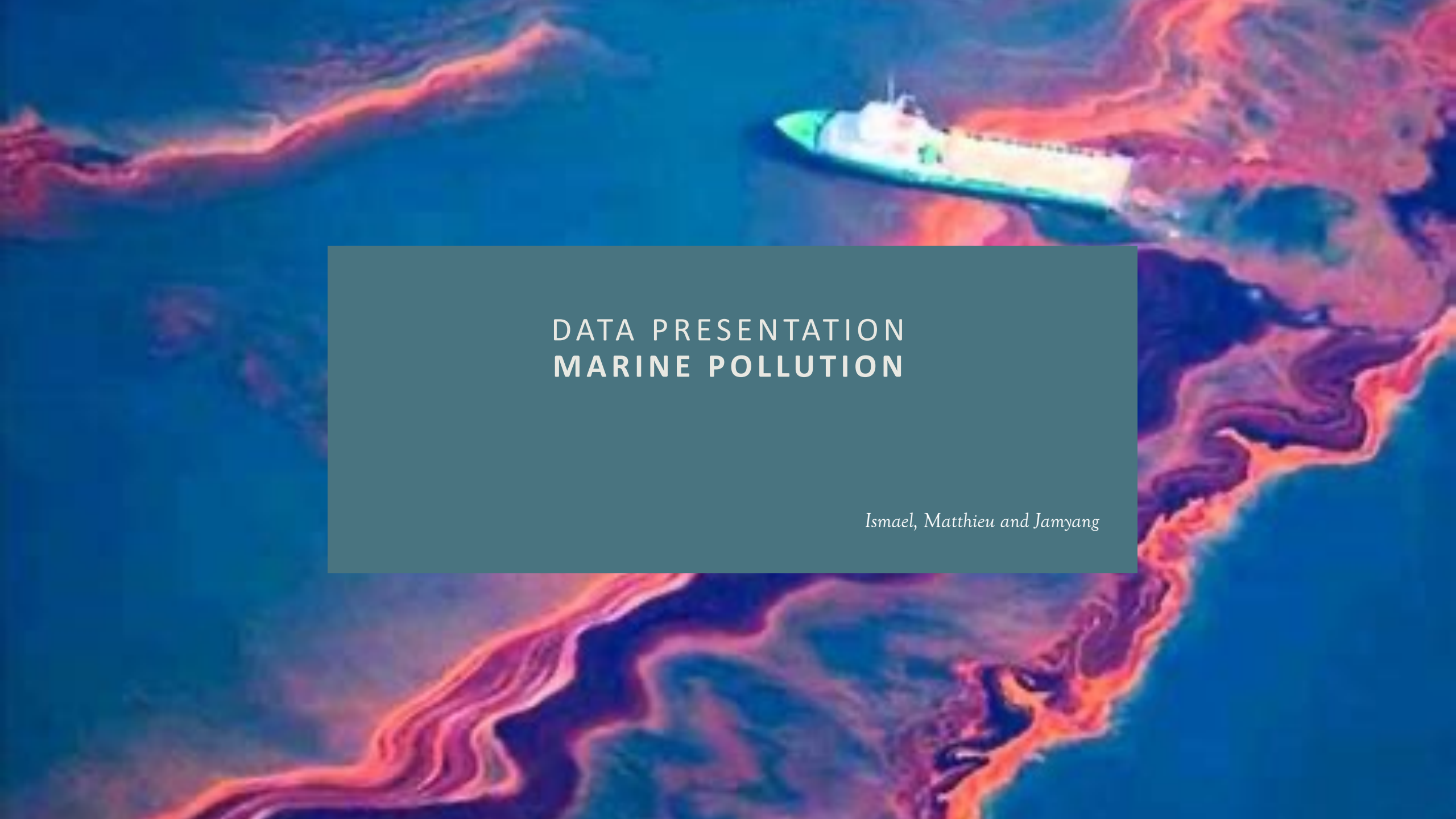


DATA PRESENTATION
MARINE POLLUTION

Ismael, Matthieu and Jamyang

DESCRIPTION OF YOUR DATASET

Marine pollution dataset from Transports and Main Roads department of Queensland Government.

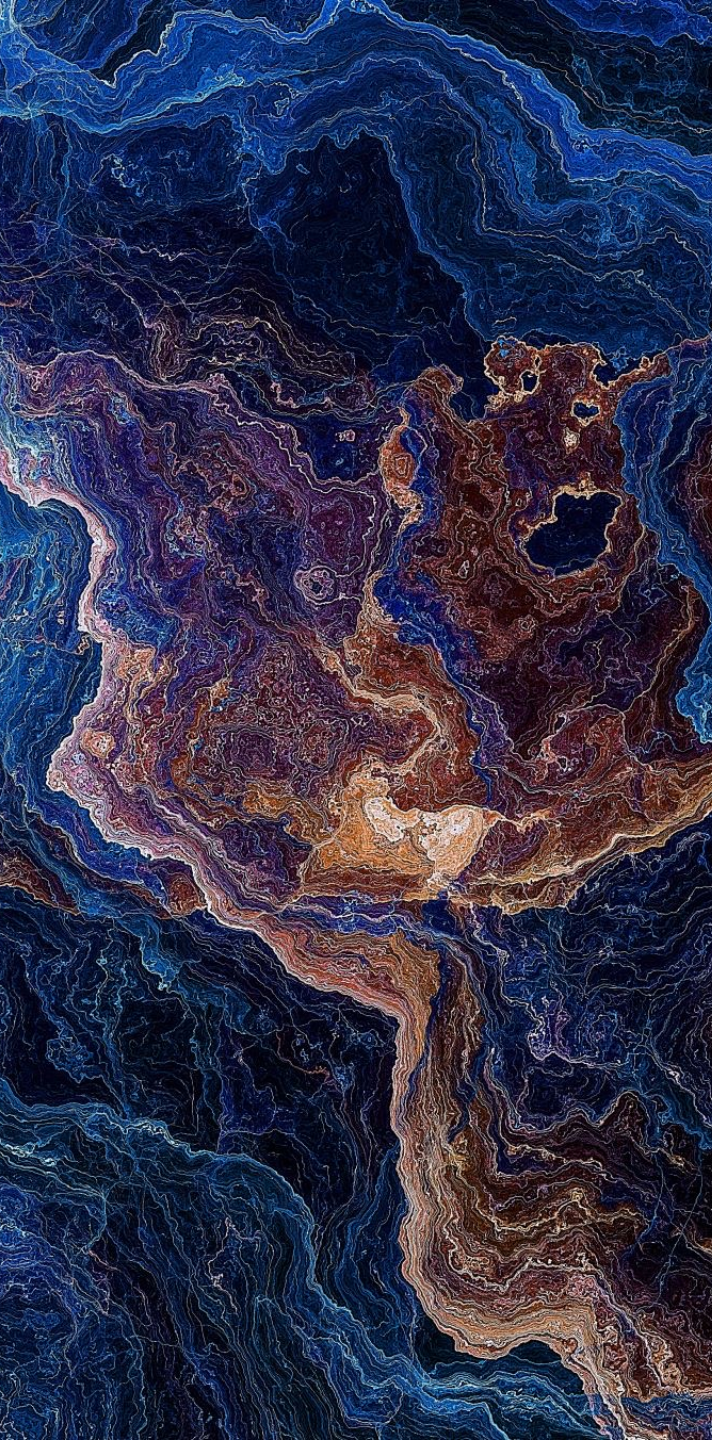
Queensland is a state situated in northeastern Australia .

File:

- 2002-2016
- 2016-2017
- 2017-2018
- 2018-2019
- 2019-2020

Source : Queensland

Queensland Government



CHALLENGES

- Understand all data
- Merge all csv files, issues with encoding
- Rename index columns to identify pollution events from 2002 to 2020
- Modify « Date » column data-type (SQL issue)
- Plenty of oil (Pollutant column)
- Different units with Estimated litres « columns » (use of Regex)

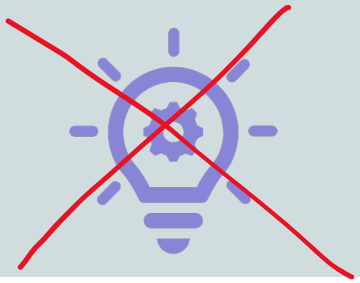
Queensland Government

PROCESS

Ease the task of cleaning : split files /person

- Split data cleaning steps (fill missing values, sort data following Queensland's Field Description, incorrect values) for each files

df_0216 : Ismael CSV (with encoding issues)
df_1617, df_1819 : Jamyang CSV
df_1718, df_pol_1920 : Matthieu (XLSX) and CSV



Abandonne the idea and compile 5 files on 1 notebook.

Merge files (check each columns from each files with describe())
Cleaning data from each column (decide together)

- Create new field (Government, others…)
- Deal with pollutant et estimated values (Challenges : Plenty of oil (Pollutant column)
- Different units with Estimated litres « columns » (use of Regex))

Clear dataframe to able to work on SQL.

- Export to SQL (add challenge)
- Modify « Date » column data-type (SQL issue)
- Highlights data with SQL

Queensland Government

IMPROVEMENTS

- Using regex
- Handle (estimated litres)
- Split tasks quickly more efficiently

.

Queensland Government

LEARNING

- Use several function in pandas (loc, concat, unique, describe)
- Data conversion between pandas and SQL

Queensland Government

COMPARISON OF THE INITIAL AND FINAL DATASETS

```
Entrée [21]: #Checking content of each column - detect missing/incorrect/duplicates values
             df_region = df_pol["Region"].unique()
             df_region

Out[21]: array(['Cairns', 'Gladstone', 'Brisbane', 'Gold Coast', 'Mackay',
                'Townsville', 'At Sea', 'Gladstone ', 'Brisbale', 'Brisbane ',
                'Cairns ', 'Hay Point', 'Bundaberg', 'Townsville '], dtype=object)

Entrée [22]: df_source = df_pol["Source"].unique()
             df_source

Out[22]: array(['ship', 'unknown', 'land', 'aircraft', 'helicopter',
                'truck in water', nan, 'coral'], dtype=object)

Entrée [23]: df_ship = df_pol["Ship Type"].unique()
             df_ship

Out[23]: array(['commercial', 'recreational', 'unknown', 'fishing', nan, 'trading',
                'tanker', 'defence', 'oil tanker', 'helicopter', 'customs',
                'n/a - museum piece', 'recreation', 'land', 'naval',
                'bulk carrier', 'trading ship', 'sailling vessel', 'navy', 'na',
                'rec', 'comm', 'fish'], dtype=object)

Entrée [24]: df_area = df_pol["Area"].unique()
             df_area

Out[24]: array(['port', 'coastal waters', 'offshore', 'gbr', 'coastal', 'gbrmp',
                'territorial sea', 'port limits', 'marina', 'inland waters'],
               dtype=object)
```

```
Diesel                          463
Sheen                           185
Bilge                           101
Other                            43
HFO                              34
                                ...
Palm acid oil                     1
Carbon from diesel engine         1
Heating                           1
 Petrol                           1
Sewage                            1
Name: Pollutant, Length: 107, dtype: int64
```

Polluants

Before

commercial	273
unknown	233
recreational	159
fishing	75
trading	44
comm	18
recreation	9
defence	9
trading ship	8
rec	8
tanker	4
oil tanker	3
customs	2
land	2
naval	2
helicopter	1
n/a - museum piece	1
bulk carrier	1
sailling vessel	1
navy	1
na	1
fish	1

After

commercial	351
unknown	243
n/a	237
recreational	177
fishing	76
government	14

Ship type

Before

coastal waters	355
port limits	348
port	313
coastal	26
gbrmp	20
gbr	15
inland waters	10
territorial sea	7
offshore	3
marina	1

After

port limits	662
coastal waters	391
great barrier reef	35
inland waters	10

HIGHLIGHTS

- Diesel is the main cause of marine pollution.

diesel	475
sheen	186
hydraulic	114
other	110
bilge	105
other oil	77
unknown	20
petroleum	11