Checkpoint II: Data Cleaning & Processing

Group: <G01>
Date: <2023/09/22>

Initial Dataset

O dataset inicial era constituído por cinco tabelas de dados, todas encontradas na internet e com os links disponíveis no relatório do CPI. Segue-se então a descrição das tabelas iniciais. Abaixo encontram-se data samples das várias tabelas. As […] significam informação irrelevante e/ou repetitiva e/ou com o mesmo significado.

(1) Country, City, AQI Value, AQI Category, CO AQI Value, CO AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI Value, NO2 AQI Category, PM2.5 AQI Value, PM2.5 AQI Category
Russian Federation, Praskoveya,51, Moderate,1, Good,36, Good,0, Good,51, Moderate

(2) Entity, Code, Year, Deaths - Chronic respiratory diseases - Sex: Both - Age: Under 5 (Rate), […]
Afghanistan, AFG,1990,5.92,52.07,3.09,12.63,94.84,818.1,183.38

(3) Entity, Code, Year, Agriculture, Land-use change and forestry, Waste, […]
Afghanistan, AFG, 1990, 8070000, -2390000, 1230000, 50000, 410000, 1670000, 170000, 80000, 280000, 0, 20000

(4) […], ParentLocationCode, ParentLocation, Location type, SpatialDimValueCode, Location, Period type, Period, IsLatestYear, Dim1 type, Dim1, Dim1ValueCode, Dim2 type, Dim2, Dim2ValueCode, Dim3 type, Dim3, Dim3ValueCode, DataSourceDimValueCode, DataSource, FactValueNumericPrefix, FactValueNumeric, FactValueUoM, FactValueNumericLowPrefix, FactValueNumericLow, FactValueNumericHighPrefix, FactValueNumericHigh, Value, […]
[…], text, AFR, Africa, Country, KEN, Kenya, Year, 2019, true, Residence Area Type, Cities, CITY, […]10.01,6. 29,13.74,10.01 [6.29 – 13.74], EN, 2022-08-11T23:00:00.000Z

(5) country_code; region_name; sub_region_name; intermediate_region; country_name; income_group; year; total_gdp; total_gdp_million; gdp_variation
ABW; Americas; Latin America and the Caribbean; Caribbean; Aruba; Ingreso alto;1960;0.0;0.0;0.0

Número	Nome do Ficheiro	(#Atributos x #Itens)
(1)	global-air-pollution-dataset.csv	12 * 23463
(2)	respiratory-disease-death-rates-by-age.csv	10 * 6840
(3)	ghg-emissions-by-sector.csv	14 * 6150
(4)	who-data.csv	34 * 9450
(5)	countries-gdp-hist.csv	10 * 13330
Tamanho Total do Dataset		890656

Selected/Derived Data

(1) + (2) + (3) + (5) {"Country":"Fiji", "Year":2016,"Age: Under 5":4,"Age: All Ages":44,"Age: 5-14 years":2, "Age: 15-49 years":12,"Age: Age-standardized":76, "Age: 70+ years":673,"Age: 50-69 years":105,"Agriculture":420000.0,"Land-use change and forestry":-2540000.0,"Waste":130000.0,"Industry":160000.0,"Manufacturing and construction":220000.0, "Transport":760000.0,"Electricity and heat":330000.0, "Buildings":120000.0, "Fugitive emissions":0.0, "Other fuel combustion":10000.0,"Aviation and shipping":310000.0, "Total emissions":-77984.0, "GDP":4930204219.0, AQI": "..", "CO AQI": "..", "Ozone AQI": "..", "NO2 AQI": "..", "PM2.5 AQI": ".."

(4) {"Continent": "Africa", "ISO3": "KEN","Country": "Kenya","Year": 2019,"Type": "Cities", "PM2.5": 10.01}

A única derived measure que achamos ser uma boa adição ao nosso dataset, foi o atributo "Total emissions"(Total emissions = sum(Agriculture,Land-use change and forestry, Waste, Industry, Manufacturing and construction, Transport, Electricity and heat, Buildings, Fugitive emissions, Other

fuel combustion,Aviation and shipping)) que foi adicionado na tabela **(2) + (3) + (5)**. Achamos que seria um bom atributo para nos ajudar a responder com mais detalhe à quarta pergunta (e assim comparar o total de emissões mais diretamente). Não achamos necessário existirem mais derived measures pois os dados restantes que temos já servem para responder às nossas questões (isto é confirmado no ùltimo tópico: Mapping).

Data Abstraction

As abstrações de dados que decidimos escolher, mas que estão suscetíveis a mudanças, são Time Series Data e Geospatial Maps, pois queremos analisar o desempenho de países com o tempo e posição geográfica.

Categorias	Tabela	Dados	Semânticas
Contínuo	**(4)** e **(1) + (2) + (3) + (5)**	Ano	Ano(data) em questão
	(1) + (2) + (3) + (5) **(2) + (3) + (4)**	Land-use change and forestry	Emissões da silvicultura em toneladas de greenhouse gases
		Total Emissions	Total de emissões em todos os setores de um país
		AQI, CO AQI, Ozone AQI, [...]	Indexes de Qualidade do Ar (0-50 good, 51-100 moderate, 101-150 Unhealthy for Sensitive Groups, 151-200 Unhealthy, 201-300 very Unhealthy, 301-higher hazarous)
Rácio	**(1) + (2) + (3) + (5)**	GDP	Produto interno bruto do país
		Age: All Ages, Age: Under 5, [...]	Morte por doença respiratória num intervalo de idade
		Industry, Transport, Buildings, [...]	Emissões das Indústrias/setores em toneladas de greenhouse gases
	(4)	PM2.5	Quantidade de Partículas finas com 2,5 micrómetros ou menos de diâmetro
Nominal	**(4)** e **(1) + (2) + (3) + (5)**	Pais / ISO3	Nome do país / código do país
	(4)	Continente	Nome do continente
		Tipo de localidade	Zona/região de um país (aldeia/vila/cidade)

Data Processing

O processamento dos dados foi feito utilizando a biblioteca de *Python*: pandas, e com ela conseguimos retirar várias colunas desnecessárias para a nossa visualização das diferentes tabelas, normalizar nomes de atributos para nomes mais simples e fáceis de entender. Apenas encontramos valores em falha ao unificar as *dataset* (deaths_emissions_gdp. json, onde usamos as chaves ano e país para os cruzar) onde optámos por preencher com valor sentinela 0.0 nos casos de mortes por doenças respiratórias e emissões de gases poluentes. Ainda, optamos por omitir países que não constavam no mapa do Geo Tutorial como *outliers* (ex: Cabo Verde, Aruba).

Mapping (Data sample/Questions)

1ª: Qual a melhor região para morar de um certo país? -> **(4)** (ISO3, Year, Type, PM2.5)
2ª: Como é que qualidade do ar afeta a taxa de mortalidade por doenças pulmonares em diferentes regiões ou cidades? -> **(1)+(2)+(3)+(5)** (Year, Country, Total Emissions, todos os atributos com Age: …)
3ª: Qual a indústria que mais danifica os nossos pulmões? -> **(1)+(2)+(3)+(5)** (Year, Country, Total Emissions, todos os atributos que representem um setor)
4ª: Que países estão a contribuir para se tornarem carbon neutral e como se comparam com o mundo? -> **(4)** (ISO3, Year, PM2.5), **(1)+(2)+(3)+(5)** (Total emissions, atributos relacionados com nomes de setores, Country, Year, todos os atributos com AQI Value)
5ª: Qual o impacto ambiental do crescimento económico? -> **(1)+(2)+(3)+(5)** (Country, Total emissions, GDP, Year), **(4)** (ISO3, PM2.5, Year)