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) {"Country": "Russian Federation", "AQI": 42.51, "CO AQI": 1.04, "Ozone AQI": 34.13, "NO2 AQI": 1.11, "PM2.5 AQI": 34.51}

(2) + (3) + (5) {"Country":"Fiji", "Year":2016.0,"Age: Under 5":3.06,"Age: All Ages":43.88,"Age: 5-14 years":1.82, "Age: 15-49 years":11.27,"Age: Age-standardized":75.29,"Age: 70+ years":672.76,"Age: 50-69 years":104.83,"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}

(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” 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 **(2) + (3) + (5)**	Ano	Ano(data) em questão
	(2) + (3) + (5)	Land-use change and forestry	Emissões da silvicultura em toneladas
Rácio	**(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
	(1)	AQI, CO AQI, Ozone AQI, [...]	Indexes de Qualidade do Ar
	(4)	PM2.5	Partículas finas com 2,5 micrómetros ou menos de diâmetro
Nominal	**(1)**, **(4)** e **(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, ao contrário de *null* nas colunas onde não calculamos as médias, medianas e quartis. Ainda, optamos por omitir países que não constavam no mapa do Geo Tutorial como *outliers* (ex: Cabo Verde, Aruba).

Mapping (Data sample/Questions)

Q1	Esta questão é respondida fazendo uso da tabela **(4)** do nosso dataset, pois possui a comparação das partículas poluentes do ar, ao longo dos anos, de várias zonas (cidade/rural/aldeia) de um país.
Q2 e Q3	Estas duas questões são respondidas através da tabela **(2) + (3) + (5)** pois esta possui as mortalidades por doenças respiratórias por idade, ano e país, e possui também as emissões ao longo dos anos, por setor e por país.
Q4	A quarta pergunta é respondida utilizando as tabelas **(1)** (para fazer a comparação das diferentes partículas nocivas para a qualidade do ar), da tabela (4) (pois possui os o valor total dos níveis de PM2.5 ao longo dos anos) e da tabela **(2) + (3) + (5)** (contem o valor total de emissões, derived measure mencionada acima, de cada país por setor ao longo dos anos).
Q5	Para esta questão, vamos fazer uso das tabelas **(2) + (3) + (5)** (contém o GDP de cada país ao longo dos anos e contem o total de emissões de cada país ao longo dos anos) e da tabela **(4)** (possui o total de partículas PM2.5 ao longo dos anos por país).