PROJECTO - CATÁSTROFES NATURAIS

Stage 1

Integração e Processamento Analítico de Informação - 2022/2023

Ana Araújo nº 59457
Francisco Vicente nº 59363
João Faia nº 47051
Tomás Oom nº 59447

Índice

Introdução

A análise de dados é uma atividade essencial em diversas áreas, nomeadamente em finanças, ciência, saúde e política, entre outras. Com a crescente geração e armazenamento de dados, a construção de um data warehouse torna-se uma estratégia fundamental para gerenciar e explorar as informações de forma eficiente e eficaz. A integração de diferentes fontes de dados num único repositório, permitindo o acesso e análise de dados de várias perspetivas e dimensões, é um dos principais benefícios de um data warehouse.

No contexto deste projeto, o objetivo é explorar os dados relacionados a catástrofes naturais e entender o seu impacto em áreas cruciais, como agricultura, desenvolvimento dos países e turismo. A escolha dessas áreas é relevante, uma vez que esses setores são altamente influenciados pelas condições climáticas e pelos eventos naturais.

Na primeira etapa do projeto, a procura e a identificação de fontes de dados relevantes para o processo de negócio escolhido foram realizadas. A análise dos dados, considerando aspetos como quantidade, a sua dimensionalidade e a sua adequabilidade, foi conduzida para selecionar um conjunto de dados rico e multidimensional. A seleção de dados global para um longo período de tempo foi realizada para proporcionar uma visão abrangente e histórica dos impactos das catástrofes naturais.

Durante a análise dos dados, foi utilizada a linguagem de programação python com a ferramenta deepnote. A representação das informações em um diagrama permitiu a visualização da conexão entre as diferentes fontes de dados e a identificação das hierarquias que fornecerão riqueza informativa para o projeto.

Um dos maiores desafios deste projeto é a complexidade dos dados relacionados às catástrofes naturais. São muitas as variáveis envolvidas, como a intensidade do evento, a localização, o tipo de desastre, entre outros. Além disso, é importante considerar a variabilidade e a heterogeneidade dos dados de diferentes fontes. Por isso, a seleção cuidadosa de fontes de dados relevantes e a análise minuciosa dos dados são fundamentais para o sucesso do projeto.

Espera-se que a construção e modelagem do data warehouse forneça uma visão mais completa e profunda sobre os impactos das catástrofes naturais na agricultura, desenvolvimento dos países e turismo. A estrutura de hierarquias identificadas oferecerá a riqueza necessária para a navegação e construção do projeto, permitindo uma análise multidimensional e uma visão integrada dos dados.

1 Descrição de fonte de dados

1.1 EM-DAT (Emergency Events Database)

EM-DAT (Emergency Events Database) é uma base de dados global de eventos de emergência, mantida pelo Centro de Investigação sobre Epidemiologia de Desastres (CRED) da Universidade Católica de Louvain, na Bélgica. A base de dados foi criada em 1988 e contém informações sobre desastres naturais desde 1900 até ao presente.

A base de dados EM-DAT é uma das mais completas e confiáveis fontes de informação sobre desastres naturais em todo o mundo. Esta fonte de dados inclui informações sobre desastres como terremotos, furacões, cheias, secas, incêndios florestais e deslizamentos de terra, entre outros. Para cada evento, a base de dados registra informações como a localização, data, tipo de desastre, número de vítimas, danos materiais e medidas tomadas para prevenir ou mitigar os efeitos do desastre.

A base de dados EM-DAT é utilizada por organizações internacionais, governos e agências de ajuda humanitária para monitorizar a evolução e as tendências dos desastres naturais em todo o mundo. Além disso, a base de dados é amplamente utilizada por investigadores e cientistas para estudar as causas e os efeitos dos desastres naturais e para desenvolver estratégias de prevenção.

O conjunto de dados EM-DAT de desastres naturais utilizado neste projeto é uma versão da base de dados EM-DAT disponibilizada no Kaggle, contendo informações sobre desastres naturais ocorridos entre 1900 e 2023. O conjunto de dados inclui informações sobre os eventos, tais como localização, data, tipo de desastre, número de mortos e danos materiais. O objectivo de utilizar este conjunto de dados serve para fazer análises e estudos de desastres naturais no mundo e estudar os seus impactos nas áreas da agricultura, desenvolvimento de países e turismo. A descrição de todas as variáveis do conjunto de dados, está especificada na tabela 1.

Tabela 1. Informação sobre os dados da dataset EM-DAT

	Coluna	Descrição	Tipo de dados
	Dis No	Identificador único (Year+Seq+ISO)	Categórico
Informação sobre a identificação da catástrofe	Disaster Group	Nível mais abrangente de classificação de desastres e divide-os em três grupos principais: desastres naturais, desastres tecnológicos e desastres complexos. Neste dataset apenas estão incluídos os desastres naturais.	Categóricos
	Disaster Subgroup	Divide os desastres naturais em sete subgrupos: climáticos, hidrológicos, meteorológicos, geofísicos, extraterrestres, biológicos e mistos.	Categóricos
	Disaster Type	Nível mais específico de classificação que identifica o tipo de desastre natural, como furacão, terremoto, inundação, seca, etc.	Texto
	Disaster Subtype	Nível mais detalhado de classificação e fornece informações adicionais sobre o tipo de desastre, como sua intensidade, duração, características específicas, entre outros.	Texto
	Event Name	Nome do evento específico que causou o desastre natural registrado.	Texto
	Entry Criteria	Indica os critérios que devem ser atendidos para que um evento seja registrado na base de dados	Texto
	Origin	Origem ou causa do desastre natural registrado	Texto
	Dis Mag Value	Refere-se à magnitude ou intensidade de um desastre	Numérico
	Dis Mag Scale	Fornece informações sobre a escala utilizada para medir a magnitude ou intensidade de um desastre na coluna "Dis Mag Value"	Texto
	Associated Dis	Evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre registrado na linha correspondente	Texto
	Associated Dis2	Outro evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre registrado na linha correspondente	Texto
Informação sobre a acção política e humanitária	OFDA Response	Refere-se às intervenções da Office of U.S. Foreign Disaster Assistance (OFDA), que é uma agência do governo dos Estados Unidos responsável por coordenar a resposta a desastres em países estrangeiros (“Yes” ou “No”)	Booleanos
	Appeal	Indica se houve um apelo de ajuda humanitária feito em relação ao desastre registrado na linha correspondente (“Yes” ou “No”)	Booleanos
	Declaration	Indica se houve uma declaração oficial de estado de emergência ou desastre feita em relação ao desastre registrado na linha correspondente (“Yes” ou “No”)	Booleanos
	Aid Contribution	Refere-se à ajuda financeira ou material fornecida por países ou organizações para ajudar no alívio ou recuperação de uma situação de desastre.	Numérico

Tabela 1. Informação sobre os dados da dataset EM-DAT (continuted)

	Coluna	Descrição	Tipo de dados
Informação sobre a localização	Country	País onde ocorreu o desastre	Texto
	ISO	Código do país onde ocorreu o desastre	Categóricos
	Region	Região onde ocorreu o desastre	Texto
	Continent	Continente onde ocorreu o desastre	Texto
	Location	Localização onde ocorreu o desastre	Texto
	Latitute	São as coordenadas geográficas onde ocorreu o desastre natural registrado	Numéricos
	Longitude	São as coordenadas geográficas onde ocorreu o desastre natural registrado	Numéricos
	Local Time	É o horário local em que o desastre ocorreu	Data e hora
	River Basin	É a bacia hidrográfica onde ocorreu o desastre natural registrado	Categóricos
Informação temporal	Year	Ano em que ocorreu a catástrofe	Numérico
	Seq	Número de sequência em que ocorreu o desastre	Numérico
	Start Year	Representa o ano em que o desastre natural registrado começou ou teve início.	Numéricos
	Start Month	Representa o mês em que o desastre natural registrado começou ou teve início.	Numéricos
	Start Day	Representa o dia em que o desastre natural registrado começou ou teve início.	Numéricos
	End Year	Representa o ano em que o desastre natural registrado acabou.	Numéricos
	End Month	Representa o mês em que o desastre natural registrado acabou.	Numéricos
	End day	Representa o dia em que o desastre natural registrado acabou.	Numéricos
Informação sobre os afetados	Total Deaths	Representa o número total de pessoas que morreram em consequência do desastre natural	Numéricos
	No injured	Número de pessoas que ficaram feridas como resultado do desastre natural	Numéricos
	No Affected	Número de pessoas afetadas pelos desastres naturais	Numéricos
	No Homeless	Indica o número de pessoas que ficaram desalojadas como resultado do desastre natural.	Numéricos
	Total Affected	Indica o número total de pessoas afetadas pelo desastre natural	Numéricos
Informação sobre os danos monetários	Reconstruction Costs ('000 US$)	Custo estimado para reconstruir as áreas afectadas pelo desastre, em milhares de dólares americanos.	Numéricos
	Insured Damages ('000 US$)	Valor estimado dos danos segurados causados pelo desastre, em milhares de dólares americanos	Numéricos
	Total Damages ('000 US$)	Custo total estimado dos danos causados pelo desastre, em milhares de dólares americanos	Numéricos
	CPI	Consumer Price Index (Índice de Preços ao Consumidor) é uma medida estatística que mede a variação dos preços de um conjunto de bens e serviços consumidos pelas famílias ao longo do tempo.	Numéricos

1.2 World Bank

O World Bank Open Data é uma plataforma online que disponibiliza gratuitamente dados e indicadores económicos, sociais e ambientais de todos os países membros. A plataforma permite acesso a vários indicadores de desenvolvimento, incluindo dados sobre a pobreza, educação, saúde, meio ambiente, finanças e comércio, entre outros.

1.2.1 World Data Bank - GDP

Este dataset refere-se ao indicador de Produto Interno Bruto (PIB) em dólares dos países, que é uma medida amplamente utilizada para medir o tamanho e a saúde económica de um país. O PIB é calculado somando o valor de todos os bens e serviços finais produzidos dentro das fronteiras de um país em um determinado período. A base de dados é mantida pelo World Bank e inclui informações anuais do PIB para diversos países, bem como regiões e grupos.

1.2.2 World Data Bank - Total Population

Este dataset contém informações sobre a população total de países n o mundo. A fonte de dados é o World data bank e a unidade de medida é o número total de habitantes. Os dados podem ser filtrados por país, região e ano. A base de dados é atualizada regularmente com novos dados à medida que se tornam disponíveis.

1.2.3 World Data Bank - Unemployment

Esta é uma base de dados do World Data Bank que fornece informações sobre a taxa de desemprego total em percentual da força de trabalho total para países no mundo. A base de dados permite a visualização dos dados e fazer uma comparação entre os países. Os dados estão disponíveis desde 1991 e são atualizados anualmente.

1.2.4 World Data Bank - Turismo

O conjunto de dados referente ao turismo foi extraído do World Data Bank e detalha o número de cidadãos internacionais que entram no país (número de chegadas). O dataset apresenta uma lacuna de dados desde 1960 até 1994, sendo estes anuais desde 1995 até 2020, para todos os países/regiões.

Tabela 2. Informação sobre as variáveis dos dataset retirados do World Data Bank

Variável	**Descrição**
Series Name	O nome da série de dados relativamente a GDP, População, Desemprego e Turismo
Series Code	O código da série de dados relativamente a GDP, População, Desemprego e Turismo
Country Name	O nome do país ou região.
Country Code	O código do país ou região.
Year	O ano em que os dados foram coletados ou relatados

1.2.5 FAOSTAT - Agricultura

A FAOSTAT é a base de dados estatísticos da Organização das Nações Unidas para Alimentação e Agricultura (FAO). Contém informações detalhadas sobre a produção, comércio e consumo de alimentos e produtos agrícolas no mundo. Os dados da FAOSTAT são de fontes nacionais e internacionais, incluindo governos, organizações não-governamentais e organizações internacionais, e abrangem uma ampla variedade de tópicos, como agricultura, pesca, segurança alimentar, nutrição, uso da terra e mudança climática.

Tabela 3. Informação sobre as variáveis do dataset AGRICULTURA

Variável	**Descrição**
Domain Code	Código do domínio a que pertence o conjunto de dados (por exemplo, "QI" para dados de qualidade e consumo de alimentos)
Domain	O nome do domínio a que pertence o conjunto de dados.
Area Code (M49)	Código numérico único para a área geográfica em que os dados foram coletados, conforme definido pelo sistema M49 das Nações Unidas.
Area	O nome da área geográfica em que os dados foram coletados.
Element Code	Código numérico único para o tipo de informação que está a ser relatada (por exemplo, "5510" para dados de uso de água em agricultura).
Element	O nome do tipo de informação que está a ser relatada.
Item Code (CPC)	Código numérico único para o tipo de produto ou recurso agrícola em que os dados estão sendo relatados, conforme definido pelo Sistema de Classificação de Produtos das Nações Unidas (CPC).
Item	O nome do produto ou recurso agrícola em que os dados estão a ser relatados.
Year Code	Código numérico único para o ano em que os dados foram adquiridos.
Year	O ano em que os dados foram adquiridos.
Unit	A unidade de medida usada para os dados relatados.
Value	O valor numérico dos dados relatados em forma de índice (2014-2016 = 100).
Flag	Um código de um ou dois caracteres que indica a qualidade ou a fonte dos dados relatados.
Flag Description	Uma descrição do código da Flag que indica a qualidade ou a fonte dos dados relatados.

2 Pré-processamento de Dados

Uma primeira análise aos datasets revela uma necessidade de processamento dos dados, de modo a que estes fiquem uniformizados. O pré-processamento de dados é uma etapa fundamental na construção de um data warehouse, especialmente quando se trata de integrar múltiplas fontes de dados. Este envolve várias etapas, incluindo a limpeza dos dados, a transformação de dados para um formato adequado para análise e a integração de dados de diferentes fontes. A limpeza dos dados envolve a remoção de valores duplicados, valores em falta (ou imputação), valores inconsistentes e outros erros que podem afetar negativamente a qualidade dos mesmos. A integração dos dados é importante para combinar informações de diferentes fontes num formato consistente e uniformizado, permitindo garantir a qualidade dos dados da warehouse e a validade das análises efetuadas.

Valores nulos

Para garantir a qualidade dos dados, é importante substituir os valores em branco e os valores considerados desconhecidos por um valor apropriado, como o NA. É essencial garantir que todas as tabelas utilizadas no processo estejam preenchidas corretamente. A abordagem da substituição dos valores nulos por "NA". Essa abordagem garante que os dados sejam mais precisos e confiáveis, proporcionando uma base sólida para análise posterior.

2.1 Dataset - EM-DAT

2.1.1 Valores em falta

Neste dataset começamos por ver os valores nulos e foram obtidos os seguintes:

Dis No	0
Year	0
Seq	0
Glide	14845
Disaster Group	0
Disaster Subgroup	0
Disaster Type	0
Disaster Subtype	3295
Disaster Subsubtype	15464
Event Name	12605
Country	0
ISO	0
Region	0
Continent	0
Location	1810
Origin	12515
Associated Dis	12994
Associated Dis2	15809
OFDA Response	14852
Appeal	14008
Declaration	13239
AID Contribution ('000 US$)	15792
Dis Mag Value	11514
Dis Mag Scale	1217
Latitude	13801
Longitude	13801
Local Time	15420
River Basin	15239
Start Year	0
Start Month	395
Start Day	3610
End Year	0

End Month	700
End Day	3529
Total Deaths	4778
No Injured	12453
No Affected	6946
No Homeless	14104
Total Affected	4488
Reconstruction Costs ('000 US$)	16534
Reconstruction Costs Adjusted ('000 US$)	16534
Insured Damages ('000 US$)	15459
Insured Damages Adjusted ('000 US$)	15459
Total Damages ('000 US$)	11199
Total Damages Adjusted ('000 US$)	11203
CPI	39

Existem várias razões pelas quais um registo de um desastre natural pode ter valores em branco no dataset de catástrofes naturais de EM-DAT:

A informação pode não estar disponível: o EM-DAT é uma fonte de dados global que depende de relatórios de várias fontes. Em alguns casos, a informação pode não estar disponível ou pode não ter sido relatada.

A existência de erros de comunicação, onde podem ser incluídos os atrasos na comunicação de informações sobre desastres naturais.

A disponibilidade limitada de dados históricos pode ser uma das razões, uma vez que muitos desastres naturais ocorreram antes do surgimento do EM-DAT ou antes que a cobertura de dados fosse tão ampla como nos dias de hoje. Portanto, é possível haver lacunas na informação histórica.

A variabilidade na qualidade dos dados no EM-DAT, que depende de muitas fontes diferentes para coletar informações sobre desastres naturais. Algumas fontes podem fornecer dados mais precisos do que outras, o que pode afectar a qualidade dos dados.

As diferentes definições de desastres naturais usadas pelas diferentes fontes, o que pode resultar em diferenças na contagem e na categorização dos eventos.

Estas podem ser algumas das causas de valores em branco no dataset EM-DAT. No entanto, é importante salientar que o EM-DAT é uma das fontes mais completas de dados sobre desastres naturais e, apesar das limitações, é amplamente utilizado para análises e estudos sobre desastres naturais em todo o mundo.

Os valores em branco da sample set foram substituídos por “NA”, à excepção do campo “OFDA”. De acordo com as diretrizes descritas na EM-DAT, para a coluna OFDA, os valores em branco foram substituídos por "No" e não por NA como nas demais colunas. Deve-se ao facto de que, inicialmente, o registo deste campo era feito apenas com a opção "Yes" e não havia registos "No". Com base nessas informações, os valores em branco na coluna OFDA foram substituídos por "No".

2.1.2 Remoção e renomeação de colunas irrelevantes

As colunas Glide, Adm Level, Admin1 Code, Admin2 Code e Geo Locations foram removidas da EM-DAT para simplificar e tornar mais uniforme o conjunto de dados. A coluna Glide não era relevante para a EM-DAT. As colunas Adm Level, Admin1 Code e Admin2 Code eram informações relacionadas à divisão administrativa dos países, que não eram úteis para a análise de dados de desastres naturais. A informação de localização geográfica também não era relevante para a análise, uma vez que a EM-DAT já fornece informações sobre a localização dos desastres naturais através das colunas Country e Location. A remoção dessas colunas permitiu que o conjunto de dados ficasse mais organizado e fácil de ser trabalhado para análises posteriores.

Para facilitar o trabalho com os dados, o nome das variáveis foi renomeado:

Tabela 4. Renomeação das variáveis do dataset EM-DAT

Nome da variável	Novo nome da variável
Dis No	DisasterID
Disaster Group	Disaster_group
Disaster Subgroup	Disaster_subgroup
Disaster Type	Disaster_type
Disaster Subtype	Disaster_subsubtype
Event Name	Event_name
Associated Dis	Associated_dis
Associated Dis2	Associated_dis2
OFDA Response	OFDA_response
AID Contribution ('000 US$)	AID_contribution
Dis Mag Value	Mag_value
Dis Mag Scale	Mag_scale
Local Time	local_time
River Basin	River_basin
Start Year	Start_year
Start Month	Start_month
Start Day	Start_day
End Year	End_year
End Month	End_month
End Day	End_day
Total Deaths	Total_deaths
No Injured	N_injured
No Affected'	N_affected
No Homeless	N_homeless
Total Affected	Total_affected
Reconstruction Costs ('000 US$)	Reconstruction_costs
Reconstruction Costs Adjusted ('000 US$)	Reconstruction_costs_adjusted
Insured Damages ('000 US$)	Insured_damages
Insured Damages Adjusted ('000 US$)	Insured_damages_adjusted
Total Damages ('000 US$)	Total_damages
Total Damages Adjusted ('000 US$)	Total_damages_adjusted

Nomes como “Antilles” não se referem a um país específico. Por essa razão, foram removidos do dataset EM-DAT para evitar ambiguidades e garantir a precisão dos dados. Além disso, a remoção de Antilles não afetou significativamente os resultados da análise, uma vez que os eventos registados em Antilles eram relativamente poucos em comparação com outros países individuais.

2.2 Dataset - AGRICULTURA

2.2.1 Valores em falta

O conjunto de dados relativos à agricultura não apresentou valores nulos, no entanto foi necessário realizar a remoção e a renomeação de algumas colunas.

2.2.2 Remoção e renomeação de colunas

Foi necessário realizar a remoção das colunas Domain code, Domain, Element code, Year code, Unit, Flag e Flag description, visto que estas não se consideraram relevantes para análise futura nem acrescentam informação extra ao dataset.

A renomeação das colunas Area, Item Code (CPC) e Area code (M49) foi efetuada de modo a uniformizar esta tabela com as restantes.

Tabela 5. Renomeação das variáveis do dataset AGRICULTURA

Nome da variável	Novo nome da variável
Area	Country
Item Code (CPC)	Item_code
Area Code (M49)	Country_code

Além da renomeação das colunas, foi realizada a uniformização dos valores da variável Area code (M49) tendo em conta que esta possuía os códigos dos países em formato de número e as restantes tabelas em formato de três caracteres. Para este efeito e tendo em conta que a coluna dos países/regiões já tinha sido uniformizada anteriormente, apenas se substituíram os valores pelas siglas presentes na tabela dos desastres naturais. Neste caso, o país “China Mainland” não possuía correspondência com nenhuma sigla por ser exclusivo do dataset da agricultura, assim foi atribuída a sigla “CHM”.

2.3 Dataset - GDP

2.3.1 Valores em falta

O conjunto de dados relativo ao produto interno bruto inclui diversos valores em falta:

```
Series Name     0
Series Code     0
Country         0
Country Code    0
1960          132
          ...
2017            9
2018            9
2019           11
2020           14
2021           21
```

No total, são 3336 valores nulos distribuídos ao longo dos anos, sendo que nos anos mais antigos este valor é superior aos anos mais recentes. Todos os valores em falta foram substituídos por NA.

2.3.2 Remoção e renomeação de colunas

As colunas series name e series code não apresentavam qualquer tipo de informação relevante para as análises futuras, tendo sido eliminadas.
Para uniformizar com os restantes datasets, a coluna “Country Name” foi renomeada para “Country”” e “Country Code” para “Country_code”:

Tabela 6. Renomeação das variáveis do dataset GDP

Nome da variável	Novo nome da variável
Country Name	Country
Country Code	Country_code

2.4 Dataset - TOTAL POPULATION

2.4.1 Valores em falta

Neste dataset verificamos que não existem valores em branco.

```
Country Name        0
Country Code        0
1960 [YR1960]       0
1961 [YR1961]       0
1962 [YR1962]       0
                  ...
2017 [YR2017]       0
2018 [YR2018]       0
2019 [YR2019]       0
2020 [YR2020]       0
2021 [YR2021]       0
Length: 64, dtype: int64
```

2.4.2 Remoção e renomeação de colunas irrelevantes

As colunas "Series Name" e "Series Code" do dataset do World Bank não contêm informações relevantes para análise ou modelagem de dados. Apenas fornecem a descrição e o código da variável correspondente, que são úteis apenas para fins de identificação e referência. Uma vez que essas informações podem ser obtidas facilmente a partir do site do World Data Bank, elas não são necessárias para análise de dados e, portanto, podem ser removidas do conjunto de dados para simplificar e reduzir o tamanho do arquivo.

A coluna “Country Name” foi renomeada para “Country” e a variável “Country Code” para “Country_code”, de forma a ser mais fácil o manuseamento dos dados:

Tabela 7. Renomeação das variáveis do dataset TOTAL POPULATION

Nome da variável	Novo nome da variável
Country Name	Country
Country Code	Country_code

2.5 Dataset - TURISMO

2.5.1 Valores em falta

Foi verificado a existência de valores nulos (“nan”) no dataset. Estes valores estão principalmente entre os anos 1960 e 1994, inclusive, e 2021.

Country Name	0
Country Code	0
Indicator Name	0
Indicator Code	0
1960	266
...	
2017	32
2018	36
2019	43
2020	134
2021	266

Existem um total de 10692 de valores nulos, que dizem respeito aos primeiros anos. As colunas em branco nos primeiros anos no dataset de turismo podem ser devidas a vários factores, incluindo falta de dados disponíveis, mudanças na metodologia de recolha de dados ou falta de interesse ou capacidade de alguns países em reportar os seus dados. Também é possível que algumas regiões geográficas não tenham sido registadas como países independentes até anos mais recentes, o que pode resultar em valores em falta para essas regiões nos anos anteriores. Além disso, alguns países podem não ter tido uma indústria de turismo significativa ou não tiveram um sistema de recolha de dados robusto nos seus primeiros anos de existência. Esses fatores podem ter contribuído para os valores nulos nas primeiras colunas do dataset. A substituição foi efetuada de forma similar aos restantes datasets, introduzindo “NA”.

2.5.2 Remoção e renomeação de colunas irrelevantes

No seguimento da análise de valores nulos no dataset, prosseguiu-se à remoção das colunas entre 1960 e 1994 e 2021, por não incluírem qualquer tipo de informação. Além do referido, a coluna Indicator name e Indicator code, foram também removidas por serem apenas uma identificação.

A coluna “Country Name” foi renomeada para “Country” e a coluna “Country Code” foi renomeada para “Country_code” de forma a ser mais fácil o manuseamento dos dados:

Tabela 8. Renomeação das variáveis do dataset TURISMO

Nome da variável	Novo nome da variável
Country Name	Country
Country Code	Country_code

2.6 Dataset - Desemprego

2.6.1 Valores em falta

Country Name 0
1991 [YR1991] 31
1992 [YR1992] 31
…
2018 [YR2018] 31
2019 [YR2019] 31
2020 [YR2020] 31
2021 [YR2021] 33
dtype: int64

Existem várias razões pelas quais um valor pode ser nulo no dataset de taxa de desemprego total do World Bank:

- Falta de dados: Em alguns países, pode haver falta de dados para um determinado ano ou período. Isso pode ocorrer porque o país não recolheu as informações necessárias ou porque os dados não foram disponibilizados pelo governo ou outras fontes.
- Metodologia de cálculo: A taxa de desemprego é calculada com uma metodologia específica, que pode variar de país para país ou ao longo do tempo. Se houver mudanças na metodologia de cálculo ou se diferentes fontes forem usadas, os dados podem não estar disponíveis para um determinado período.
- Erros de entrada de dados: É possível que os dados tenham sido inseridos incorretamente ou que tenham ocorrido erros durante a transmissão ou processamento dos dados.
- Diferenças culturais: As definições de desemprego podem variar entre países, o que pode levar a diferenças nos dados reportados. Alguns países podem ter uma definição mais ampla de desemprego, enquanto outros podem ter uma definição mais restrita.

Os dados em branco foram alterados para “NA”.

2.6.2 Remoção e renomeação de colunas irrelevantes

As colunas removidas deste conjunto de dados foram a Series Name e Series Code e os anos de 1960 a 1990, devido à alta quantidade de valores “NA”. De forma semelhante aos outros datasets, Country Name foi renomeado para Country, Country Code para Country_code e foram retirados os códigos dos anos.

Tabela 9. Renomeação das variáveis do dataset DESEMPREGO

Nome da variável	Novo nome da variável
Country Name	Country
Country Code	Country_code
1960 [YR1960]	1960

2.7 Uniformização dos dados

O conjunto de dados dos desastres naturais possui duas variáveis que conectam com as restantes tabelas, country e country_code, sendo necessário verificar a concordância entre estes em todas as tabelas onde ocorrem de modo a uniformizar o data warehouse. Após uma profunda análise dos países presentes em todos os datasets, concluiu-se que existem países que não estão presentes em alguns datasets mas estão presentes noutros e vice-versa. Estes países são, na sua maioria, colónias ou arquipélagos que não são reconhecidos como países independentes em todas as fontes de dados. Para realizar a uniformização dos dados, foi necessário avaliar quais as diferenças nos nomes dos países presentes em todos os datasets e trocar em todas as tabelas de modo a torná-los comparáveis.

As tabelas de dados foram extraídas da mesma fonte (World Data) e, por essa razão, foi-se verificar se as colunas que contêm nomes dos países têm valores idênticos, para que possam ser comparadas posteriormente. Assim, foi também realizada a uniformização dos valores errados num dos datasets e renomeou-se alguns países noutros datasets e comparou-se as colunas com os nomes dos países em diferentes datasets de forma a verificar se têm valores correspondentes.

Assim, criou-se uma lista de países exclusivos (sem repetições) para cada dataframe de dados (EM-DAT, AGRICULTURA, GDP, POPULAÇÃO, TURISMO e DESEMPREGO). Em seguida, verificou-se se os países em do dataset GDP são iguais aos países noutros datasets. Foi criada duas novas listas chamadas "diff1" e "diff2", que contêm países que estão presentes no dataset GDP e não no dataset TURISMO e vice-versa.

Depois, foi uniformizado alguns valores errados na coluna "Country" do dataset TURISMO. Em seguida, verificou-se os nomes dos países do dataset GDP correspondem aos nomes dos países nos outros datasets. Foi criado novamente duas listas "diff1" e "diff2", que agora contêm países que estão presentes no dataset GDP e não no dataset EM-DAT e vice-versa.

Por fim, foi renomeado alguns países na coluna "Country" do dataset EM-DAT. Algumas dessas alterações incluíram a remoção de erros de digitação, a mudança nomes de países para que correspondam aos nomes do dataset GDP e a alteração do nome da antiga Tchecoslováquia para o nome atual dos seus estados sucessores.

3 Diagrama dos Datasets

De modo a tornar mais evidente as relações entre cada dataset e facilitar o processo de criação da data warehouse, foi criado um diagrama com todos os conjuntos de dados onde se conectam as variáveis partilhadas pelos diversos datasets.

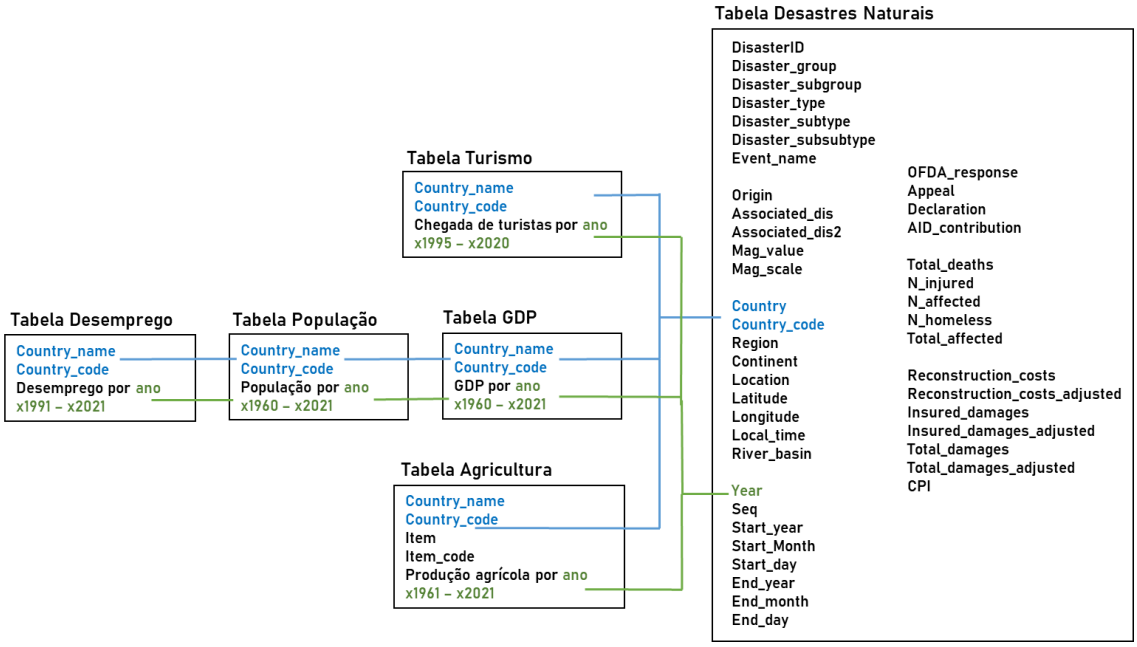

Figura 1. Diagrama dos datasets

4 Processo de Negócio

As catástrofes naturais têm um grande impacto na sociedade e no ambiente. Estas ocorrências podem ter consequências devastadoras, tais como a perda de vidas, destruição de propriedades e infraestruturas, impacto na economia e no desenvolvimento dos países. A fim de compreender e gerir melhor essas situações, é importante registar e monitorizar esses eventos.

Um dos principais objetivos de manter um registo de catástrofes naturais é estabelecer padrões que possam ajudar a prevenir ou mitigar os impactos desses eventos no futuro. Com a ajuda de dados recolhidos de diferentes fontes, podemos analisar e identificar tendências e padrões relacionados com o tipo de desastres, a sua localização geográfica e a sua intensidade. Estas informações são valiosas para tomar decisões estratégicas e planeamento a longo prazo para prevenir ou minimizar o impacto de futuras catástrofes naturais.

Por exemplo, os dados sobre catástrofes naturais podem ser utilizados para identificar áreas que são mais propensas a serem afetadas por eventos climáticos extremos, ajudando assim os governos a adotar medidas preventivas. Também podem ser utilizados para avaliar a viabilidade de um determinado setor, como o turismo ou a agricultura, em uma determinada região ou país, tendo em conta a probabilidade de ocorrerem eventos naturais adversos.

No entanto, é importante ter em conta a dimensionalidade dos dados dos datasets utilizados. Os dados sobre catástrofes naturais podem ser muito complexos e podem incluir informações de várias fontes diferentes. A interpretação desses dados pode ser difícil e requer um bom conhecimento do domínio. Além disso, é importante garantir a qualidade dos dados, pois informações imprecisas ou incompletas podem levar a tomadas de decisão erradas. Adicionalmente, neste dataset existe muita falta de informação para dados antes de 1960. A inexistência de informação para este período para levar a um estudo diferente e, consequentemente, a uma tomada de decisão inadequada.

Em resumo, a compilação de dados sobre catástrofes naturais é fundamental para entender melhor os eventos adversos e o seu impacto na sociedade e no meio ambiente. A análise desses dados pode levar a decisões mais informadas e estratégicas, que podem ter um impacto significativo na prevenção e mitigação de eventos naturais adversos no futuro. É crucial garantir a qualidade dos dados e estar ciente da complexidade dos mesmos, para que possam ser interpretados corretamente e utilizados para tomadas de decisão eficazes.

5 Questões analíticas

No seguimento dos objetivos do projecto proposto, foram elaboradas três perguntas analíticas que nos ajudem a estudar o impacto que as catástrofes naturais podem ter na agricultura, turismo e no desenvolvimento económico de um país:

1) Como as catástrofes naturais afetam a produção agrícola num determinado país e qual é o impacto na sua produção? Que tipo de desastre naturais são mais propensos a afectar a agricultura e quais os tipos de matérias primas mais afectadas?
2) Existe alguma relação entre a afluência de turistas de um país e a frequência de ocorrência de catástrofes naturais? Se sim, como essas catástrofes afetam o desenvolvimento do setor turístico e a economia do país? Quais os tipos de desastre naturais que influenciam mais o turismo?
3) Qual é o impacto das catástrofes naturais no índice de desenvolvimento de crescimento de um país e como fica afetada a taxa de desemprego e a produtividade da população? A inflação tem um valor de expressão maior nos países não desenvolvidos?