PROJECTO - CATÁSTROFES NATURAIS

Etapa 1 e Etapa 2

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 perspectivas 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[1], 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.	Categóricos
	Disaster Subtype	Nível mais detalhado de classificação e fornece informações adicionais sobre o tipo de desastre, como sua intensidade, duração, características específicas, entre outros.	Categóricos
	Event Name	Nome do evento específico que causou o desastre natural registrado.	Categóricos
	Entry Criteria	Indica os critérios que devem ser atendidos para que um evento seja registrado na base de dados	Categóricos
	Origin	Origem ou causa do desastre natural registrado	Categóricos
	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"	Categóricos
	Associated Dis	Evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre registrado na linha correspondente	Categóricos
	Associated Dis2	Outro evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre registrado na linha correspondente	Categóricos

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

	Coluna	Descrição	Tipo de dados
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
Informação sobre a localização	Country	País onde ocorreu o desastre	Categóricos
	ISO	Código do país onde ocorreu o desastre	Categóricos
	Region	Região onde ocorreu o desastre	Categóricos
	Continent	Continente onde ocorreu o desastre	Categóricos
	Location	Localização onde ocorreu o desastre	Categóricos
	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

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

	Coluna	Descrição	Tipo de dados
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[2] 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[3] 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[4] 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[5] 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)[6]. 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 - DESASTRES NATURAIS

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

Tabela 4. Renomeação das variáveis do dataset EM-DAT (continua)

Nome da variável	Novo nome da variável
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.

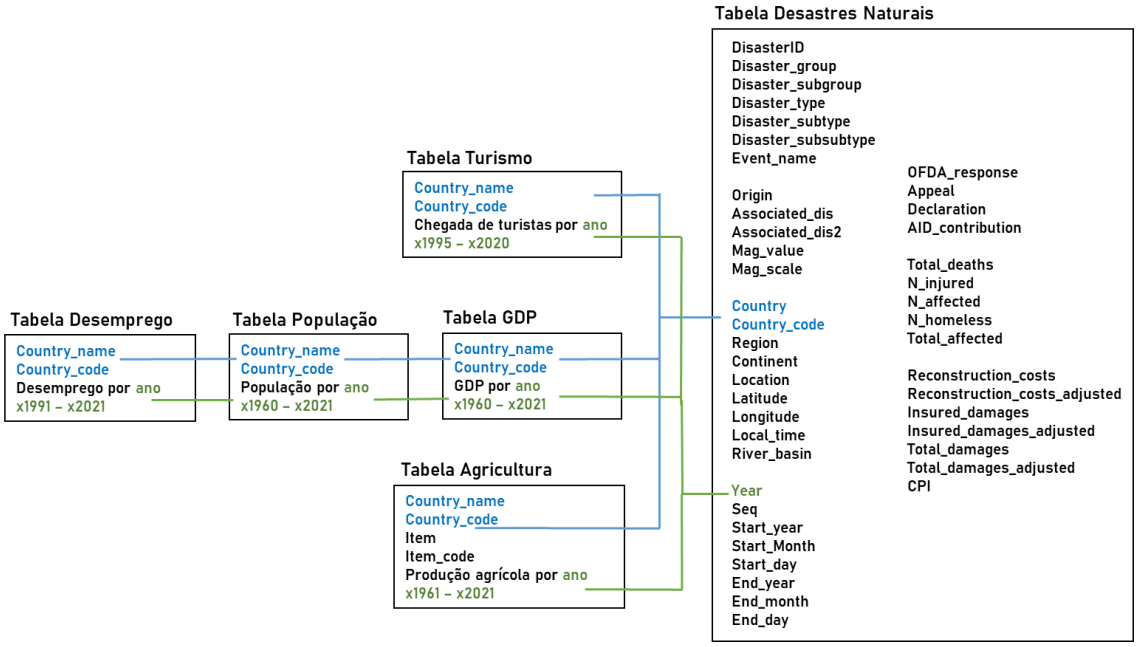

Figura 1. Diagrama dos datasets

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 projeto 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 afetar a agricultura e quais os tipos de matérias-primas mais afetadas?
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?

6 Modelação Dimensional

A modelação dimensional é uma técnica utilizada no armazenamento de dados que organiza os dados em dimensões e medidas, criando uma representação multidimensional dos dados. O objetivo da modelação dimensional é facilitar aos utilizadores finais a análise e compreensão de dados complexos, fornecendo uma estrutura que reflita os processos empresariais e a forma como os utilizadores pensam sobre os dados. A modelação dimensional baseia-se no conceito de um cubo de dados, que representa a natureza multidimensional dos dados.

Na modelação dimensional, os dados são organizados em dois tipos de tabelas: tabelas de factos e tabelas de dimensões. As tabelas de factos contêm dados quantitativos aditivos, tais como o número de mortes provocado por cada desastre ou o custo de reconstrução gasto pelo país, enquanto as tabelas de dimensões fornecem o contexto para os dados na tabela de factos, tais como a data ou a localização desses desastres. Ao combinar tabelas de factos e tabelas de dimensões, é possível criar um modelo de dados flexível e poderoso que pode suportar uma vasta gama de consultas analíticas. As tabelas de dimensões conectam-se à tabela de factos através de chaves estrangeiras, permitindo a cada facto da tabela de factos (com a sua própria chave primária) estar conectado às diversas dimensões. Deste modo, a tabela de factos juntamente com as tabelas de dimensões, permitem perceber quais as características relacionadas com cada desastre natural e analisar as diversas medidas de forma integrada.

As linhas da tabela de factos incluem o seu próprio ID, ou uma chave primária, que identifica unicamente um desastre natural, conectada às diversas dimensões pelas suas chaves substitutas. Primeiramente, a tabela de factos está associada à dimensão Tipo onde explicita o tipo de desastre natural, seguida da dimensão Data onde está incluída a informação no tempo do acontecimento (esta não está conectada diretamente à tabela de factos, pois existe uma vista SQL com os atributos da data de início do desastre e a data de fim do mesmo). Além das referidas, a dimensão Localização também é de extrema relevância com detalhes importantes sobre o país ou região onde se desenrolou o desastre e, adicionalmente, com informação relativa ao rendimento do país, o nível de turismo ou o seu desempenho na agricultura. Por fim, a última ligação da tabela de factos é com a dimensão Evento, onde inclui diversas características que descrevem o evento como a sua origem (se deriva de outro desastre), se tem desastres associados (terramoto gerar um *tsunami*) ou se o país recebeu ajuda após o desastre.

6.1 Tipo de tabela de factos e grão

A tabela de factos é do tipo transacional, uma vez que capta, em cada linha, uma transação do processo de negócio, sendo possível calcular a sua frequência e por vezes a duração de transações. Deste modo, é possível identificar o grão da tabela de factos como cada linha, ou cada transação, da tabela de factos. O grão permite identificar o nível de detalhe apresentado na linha da tabela de factos, sendo este diretamente associado ao número de dimensões conectadas e, por conseguinte, associadas ao número de chaves estrangeiras presentes nos atributos da tabela de factos.

Com base na informação anterior, a tabela de factos (tabela 10), factDisaster, irá incluir 4 dimensões com atributos filtrados de acordo com as necessidades do processo de negócio e a capacidade da *data warehouse*. Estas são a dimType, a dimDate, a dimLocation e a dimEvent. As medidas presentes na tabela de factos são essenciais para responder às questões propostas na etapa anterior e futuras questões que hipotéticos decisores possam necessitar de responder. Estas incluem as mortes provocadas pelo desastre (Total_deaths), o número de feridos (N_injured), o número de afetados (N_affected), o número de desalojados (N_homeless), a contribuição monetária de outros países (AID_contribution), o custo de reconstrução, tanto com e sem a inflação considerada (Reconstruction_costs e Reconstruction_costs_adjusted), o custo em danos cobertos pelos seguros, tanto com e sem a inflação considerada (Insured_damages e Insured_damages_adjusted) e por último o custo total de danos (Total_damages). A tabela de factos possui uma dimensão de 16568 linhas, sendo este o número de desastres registados. Tendo em conta que, teoricamente, a tabela de factos deveria conter 90% do tamanho armazenado da data warehouse, é possível prever, desde já, que não será o caso desta estratégia de modelação, existindo dimensões com maior número de linhas. Outra questão relevante para esta etapa do projeto e que terá de ser resolvida na próxima etapa é a presença de valores nulos na tabela de factos (ver Figuras 2 e 3). Tendo em conta que são de uma magnitude bastante elevada, a imputação de valores não fazia sentido. Além disso, não seria benéfico retirarmos as medidas por completo só por apresentarem bastantes valores nulos, visto que são variáveis que fazem sentido para caracterizar os desastres ocorridos. Por exemplo, um desastre natural do tipo "Drought" não vai ter qualquer tipo de impacto nas infraestruturas do país, ao contrário de um desastre do tipo "Tornado".

Tabela 10. Tabela de factos factDisaster - Descrição dos campos de dados: Identificadores, atributos e medidas aditivas.

Campo	Descrição dos dados	Origem dos dados	Valores
DisasterID	Chave Primária (Identificador do desastre natural)	Tabela **DESASTRES NATURAIS** (ID gerado sequencialmente)	N/AP
TypeKEY	Chave estrangeira (dimensão Tipo)	N/AP	N/AP
LocationKEY	Chave estrangeira (dimensão Localização)	N/AP	N/AP
StartDateKEY	Chave estrangeira (dimensão Data)	N/AP	N/AP
EndDateKEY	Chave estrangeira (dimensão Data)	N/AP	N/AP
EventKEY	Chave estrangeira (dimensão Evento)	N/AP	N/AP
Total_deaths	Número total de pessoas que morreram em consequência do desastre natural	Tabela **DESASTRES NATURAIS**	1-3700000 Média: 2758.87

Tabela 10. Tabela de factos factDisaster - Descrição dos campos de dados: Identificadores, atributos e medidas aditivas. (continua)

Campo	Descrição dos dados	Origem dos dados	Valores
N_injured	Número de pessoas que ficaram feridas como resultado do desastre natural	Tabela **DESASTRES NATURAIS**	1-1800000 Média: 2559.52
N_affected	Número de pessoas afetadas pelos desastres naturais	Tabela **DESASTRES NATURAIS**	1-330000000 Média: 873742.22
N_homeless	Número de pessoas que ficaram desalojadas como resultado do desastre natural.	Tabela **DESASTRES NATURAIS**	3-15850000 Média: 72361.08
Total_affected	Número total de pessoas afetadas pelo desastre natural	Tabela **DESASTRES NATURAIS**	1-330000000 Média: 711587.56
AID_contribution	Ajuda financeira por países ou organizações para ajudar no alívio ou recuperação de uma situação de desastre.	Tabela **DESASTRES NATURAIS**	1.0-3518530.0 Média: 20039.59
Reconstruction_costs	Custo estimado para reconstruir as áreas afetadas pelo desastre	Tabela **DESASTRES NATURAIS**	84.0-25000000.0 Média: 2543789.88
Reconstruction_costs_adjusted	custo estimado de reconstrução dos danos causados pelo desastre, ajustado à inflação e à variação cambial	Tabela **DESASTRES NATURAIS**	126.0-43922383.0 Média: 3700856.44
Insured_damage	Valor estimado dos danos segurados causados pelo desastre	Tabela **DESASTRES NATURAIS**	34.0-60000000.0 Média: 909690.38
Insured_damages_adjusted	Valor dos danos causados pelo desastre que foram segurados pelas vítimas ou por empresas, ajustado à inflação e à variação cambial para refletir os valores atuais.	Tabela **DESASTRES NATURAIS**	46.0-89913156.0 Média: 1248945.49
Total_damages	Custo total estimado dos danos causados pelo desastre	Tabela **DESASTRES NATURAIS**	2.0-210000000.0 Média: 780384.98

Tamanho: 16568 linhas

Tabela 11. Amostra exemplificativa da tabela de factos factDisaster.

DisasterID object 1900-9002-... 0% 1900-9001-I... 0% 16566 others 100%	TypeKey int64 0 - 53	LocationKey int64 0 - 6145	StartDateKey int... 0 - 9814	Reconstruction_...	Insured_damages	Insured_damag...	Total_damages f...
1900-9002-CPV	0	0	0	nan	nan	nan	nan
1900-9001-IND	0	1	0	nan	nan	nan	nan
1902-0012-GTM	1	2	1	nan	nan	nan	25000.0
1902-0003-GTM	2	2	2	nan	nan	nan	nan
1902-0010-GTM	2	2	3	nan	nan	nan	nan
...							
2023-0110-ZMB	27	6137	9807	nan	nan	nan	nan
2023-0068-ZMB	6	6137	9813	nan	nan	nan	nan
2023-0095-ZWE	4	6144	9806	nan	nan	nan	nan
2023-0022-SRB	10	6145	9814	nan	nan	nan	nan

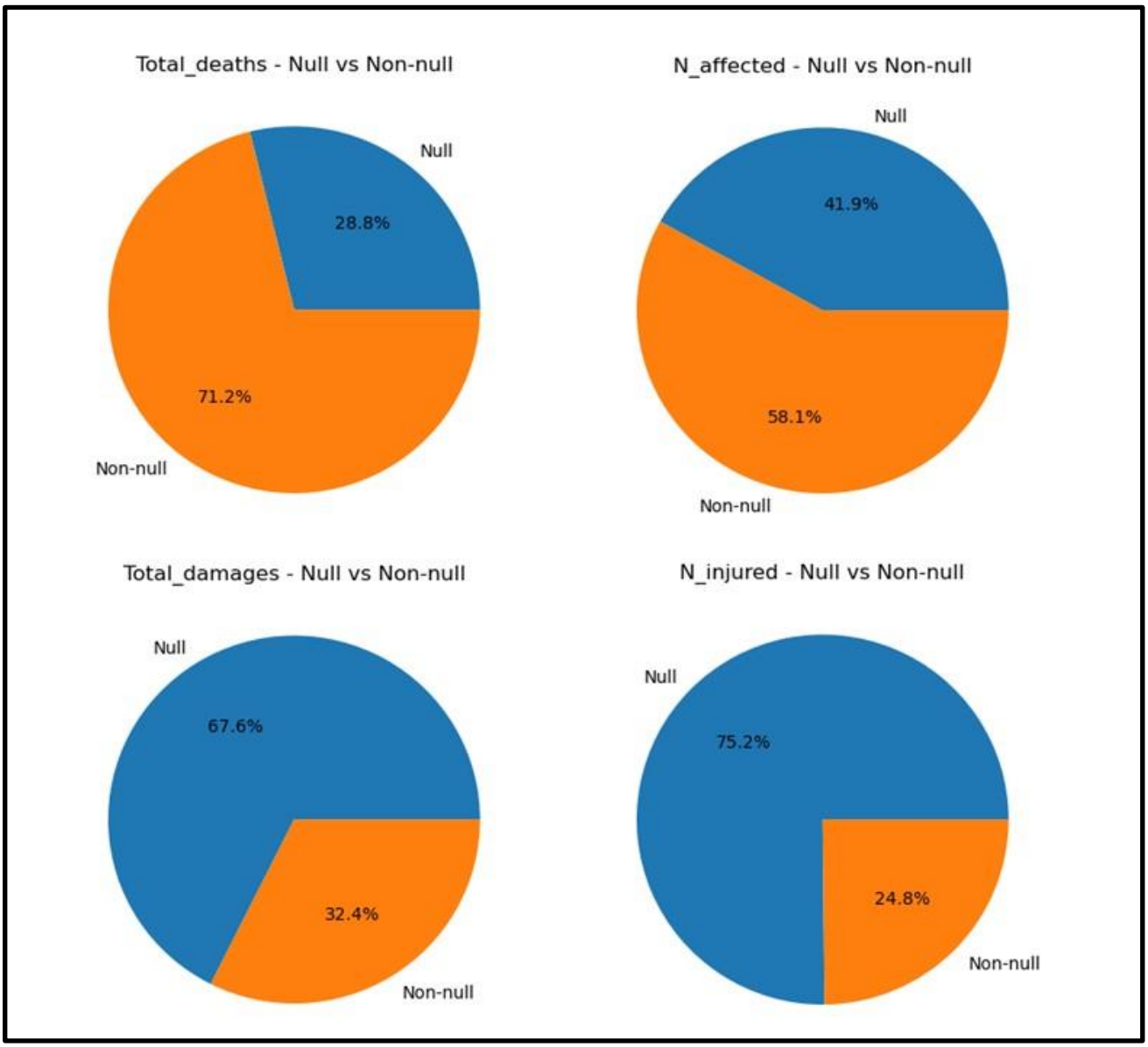

Figura 2. Gráficos de valores nulos para “Total_deaths”, “N_affected”, “Total_damage” e “N_injured” da tabela factos factDisaster.

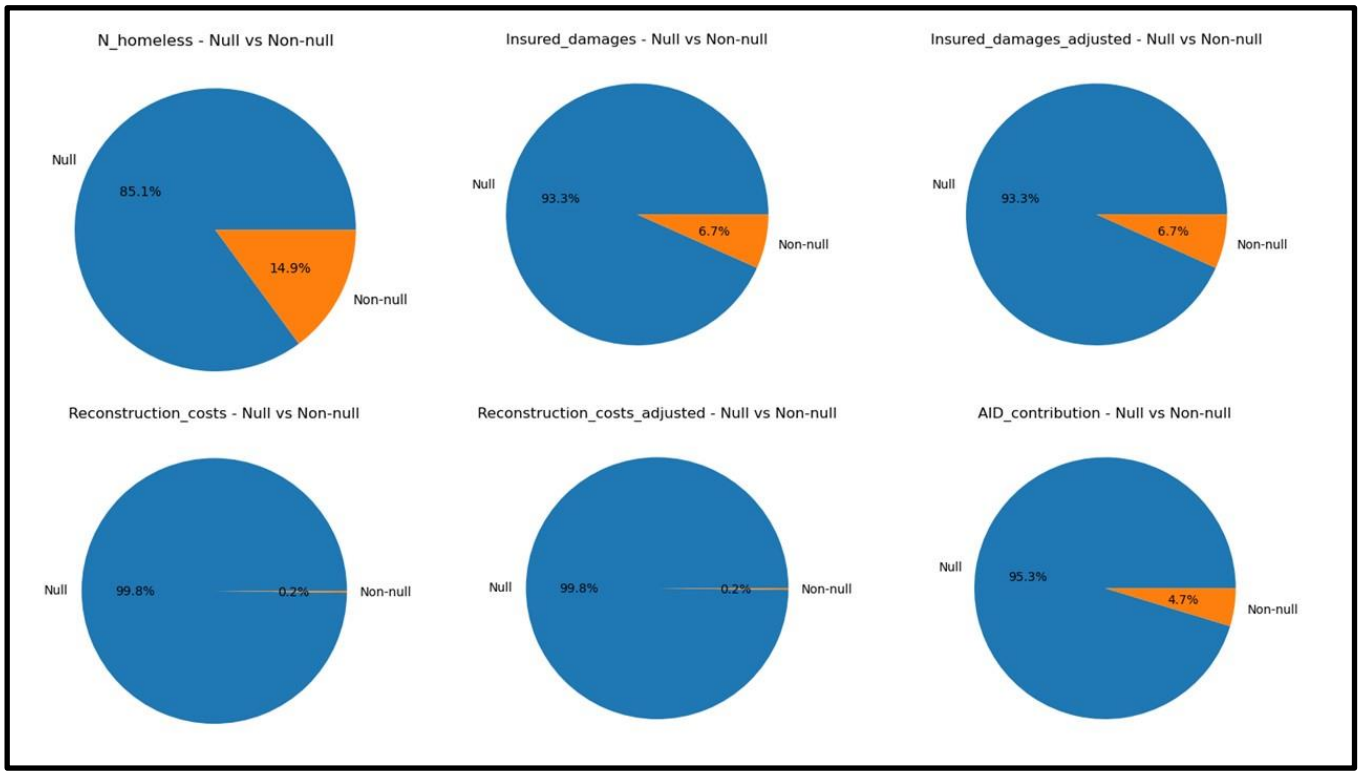

Figura 3. Gráficos circulares de valores nulos para “N_homeless”, “N_damaged”, “Insured_damage_adjusted”, “Reconstruction_costs”, “Reconstruction_costs_adjusted” e “AID_contribution” da tabela factos factDisaster.

6.2 Tabelas de Dimensões

A criação das dimensões são um passo essencial na modelação, tendo em conta que é isso que define o nível de detalhe refletido na tabela de factos. Como referido anteriormente, foram criadas quatro dimensões com características diferentes que permitem caracterizar cada linha da tabela de factos. O modelo de dados é completamente dependente desta fase do projeto, sendo de seguida caracterizado cada dimensão de modo a elucidar o seu papel e as suas características no data warehouse.

6.2.1 Dimensão Tipo - dimType

A dimensão tipo tem como objetivo a caracterização do desastre da tabela de factos identificando o grupo à qual pertence (Meteorológico, geofísico, etc.), o tipo dentro desse grupo (Tempestade, atividade vulcânica, etc.), o subtipo e o "subsubtipo". Foi gerado um ID, TypeKey, como chave substituta para identificar cada linha da tabela e irá ser usada pela tabela de factos para identificar cada tipo de desastre. Originalmente no dataset dos desastres naturais, existia uma variável que identificava o grupo dos desastres (Disaster_group), no entanto, esta não continha nenhuma informação relevante, uma vez que apenas continha a categoria "Natural disaster". Deste modo, esta variável não foi incluída na dimensão tipo, mas, tendo em conta que a nomenclatura seria bastante útil para tornar a dimensão mais compreensível, decidiu-se renomear a variável Disaster_subgroup para Disaster_group e manter as Disaster_type intactas. Assim, a dimensão tipo originou uma tabela com quatro atributos: Disaster_group, Disaster_type, Disaster_subtype e Disaster_subsubtype.

É possível compreender por observação direta dos atributos que existe uma hierarquia associada onde vai do mais abrangente, Disaster_group até ao mais detalhado, Disaster_subsubtype. Este tipo de hierarquias permitem realizar as operações de drill-down e roll-up permitindo ao utilizador escolher o nível de detalhe pretendido.

A dimensão Tipo originou uma tabela com 54 linhas e com diversos valores nulos, sendo estes mais abrangentes no último nível da hierarquia, Disaster_subsubtype. Isto deve-se ao facto de este atributo ser muito específico e a maior parte dos desastres apenas serem caracterizados por um grupo, tipo e subtipo.

Tabela 12. Tabela de dimensão Tipo - dimType. Descrição dos campos de dados: Identificadores, atributos e exemplos de valores.

Campo	Descrição dos dados	Origem dos dados	Valores
TypeKey	Chave Substituta	ID gerado sequencialmente	1-54
Disaster_group	Classificação mais abrangente de desastres naturais	Tabela **DESASTRES NATURAIS**	Ex: Meteorological
Disaster_type	Identifica o tipo de desastre natural	Tabela **DESASTRES NATURAIS**	Ex: Storm
Disaster_subtype	Nível mais detalhado de classificação de desastre	Tabela **DESASTRES NATURAIS**	Ex: Convective storm NA: 12
Disaster_subsubtype	Nível mais detalhado de classificação de desastre	Tabela **DESASTRES NATURAIS**	Ex: Tornado NA: 42

Hierarquias: Disaster_group > Disaster_type > Disaster_subtype > Disaster_subsubtype

Tamanho: 54 linhas

Tabela 13. Amostra exemplificativa da tabela de dimensão Tipo - dimType

TypeKey int64 0 - 53	Disaster_group Meteorologi... 31.5% Hydrological ... 20.4% 4 others 48.1%	Disaster_type Storm 22.2% Landslide 13% 13 others 64.8%	Disaster_subtype Not Applica... 22.2% Convective ... 16.7% 27 others 61.1%	Disaster_subsu... Not Applica... 77.8% Winter storm... 3.7% 9 others 18.5%
0	Climatological	Drought	Drought	Not Applicable
1	Geophysical	Earthquake	Ground movement	Not Applicable
2	Geophysical	Volcanic activity	Ash fall	Not Applicable
3	Geophysical	Mass movement (dry)	Rockfall	Not Applicable
4	Meteorological	Storm	Tropical cyclone	Not Applicable
...				
50	Meteorological	Storm	Convective storm	Derecho
51	Geophysical	Volcanic activity	Pyroclastic flow	Not Applicable
52	Climatological	Drought	Not Applicable	Not Applicable
53	Climatological	Glacial lake outburst	Not Applicable	Not Applicable

6.2.2 Dimensão Data - dimDate

A dimensão data tem um papel crucial na maioria dos data warehouses, sendo esta que determina todas as características temporais associadas a uma transação. Na modelação apresentada neste projeto, a dimensão data define temporalmente um desastre natural através dos seus atributos, tendo sido gerada uma chave substituta para ser associada a cada linha da tabela de factos (DateKey). Os atributos presentes na dimensão foram gerados tendo por base a informação temporal presente no dataset dos desastres naturais, no entanto, diversos novos atributos foram gerados a partir dos existentes para criar alguma redundância e oferecer uma maior flexibilidade ao utilizador, enriquecendo o presente modelo. Os atributos estabelecidos na dimensão e retirados diretamente do dataset foram o ano, o mês, o dia e o horário local (continham a hora e os minutos do desastre). A partir destes, foram gerados o nome do mês, a década (tendo em conta que se trata de desastres e alguns podem acontecer raramente), o semestre, ou trimestre, o dia da semana (número), o dia da semana (nome), as horas, os minutos e a estação do ano (importante para desastres como a seca ou tempestades). A análise destes atributos permite identificar duas hierarquias que também oferecem alguma flexibilidade ao utilizador. A primeira e de maior extensão começa no atributo mais geral, neste caso, a década e passa pelo ano, semestre, trimestre, mês (com o nome do mês associado), dia (com o número e nome do dia da semana associado), horas e minutos. A segunda hierarquia começa na década, passa pelo ano e acaba na estação do ano.

Tendo em conta a natureza dos desastres naturais, onde alguns podem ser momentâneos como um terramoto e outros podem durar dias, meses ou mesmo anos como uma pandemia, existe, não só, a necessidade de indicar a data de início do desastre, mas também, a data do fim do desastre. O facto de não ser possível utilizar duas chaves para uma mesma dimensão, obriga à criação de uma vista SQL, onde, numa fase futura, a dimensão data irá estar associada à dimStartDate e dimEndDate. Estas duas novas vistas vão estar conectadas à tabela de factos pelas chaves StartDateKey e EndDateKey, permitindo a existência de duas datas para o mesmo desastre. O facto de se

utilizar a data em dois contextos diferentes é um caso de role-playing, onde a data tem o papel de início e fim de um desastre.

A dimensão data originou uma tabela com 19221 linhas que acabou por se estender para além da tabela de factos. Esta situação deriva do caso de role-playing descrito anteriormente, onde o facto dos desastres terem duas datas associadas fez com que existissem mais combinações a ter em conta na dimensão data. A dimensão só tem mais 2653 linhas do que a tabela de factos, não sendo um valor exageradamente grande e, por isso, não é considerada uma dimensão monstro. Para além do número de linhas, a presença de valores nulos na dimensão data é uma situação indesejável, mas inevitável no contexto deste problema. Neste caso, tendo novamente em conta a natureza dos desastres naturais, é plausível que existam valores nulos pelo simples facto que alguns desastres não possuem minutos, horas, dias ou meses pois são desastres graduais como a seca. Nestes casos, também não faria sentido imputar os valores, tendo os valores nulos sido substituídos por “Unknown”.

Tabela 14. Tabela de dimensão Data - dimDate. Descrição dos campos de dados: Identificadores, atributos e medidas aditivas.

Campo	Descrição dos dados	Origem dos dados	Valores
DateKey	Chave Substituta	ID gerado sequencialmente	1-19220
Year	Número do ano	Tabela **DESASTRES NATURAIS**	1900-2023
Month	Número do mês	Tabela **DESASTRES NATURAIS**	1-12 NA: 193
MonthName	Nome do mês	Dado gerado com recurso a linguagem de programação	Ex: January NA: 193
Day	Dia do mês	Tabela **DESASTRES NATURAIS**	Ex: 1 NA: 1850
Decade	Década onde ocorreu o desastre	Dado gerado com recurso a linguagem de programação	Ex: 1980s
Semester	Número do semestre no respectivo ano	Dado gerado com recurso a linguagem de programação	Ex: 1st Semester NA: 193
Season	Estação do ano	Dado gerado com recurso a linguagem de programação	Ex: Winter NA: 1851
Quarter	Número do trimestre no respectivo ano	Dado gerado com recurso a linguagem de programação	Ex: 2nd Quarter NA: 193
Weekday	Época sazonal	Dado gerado com recurso a linguagem de programação	1-7 NA: 1851
WeekDayName	Nome dia da semana	Dado gerado com recurso a linguagem de programação	Ex: Monday NA: 1851
LocalTime	Hora do incidente	Dado gerado com recurso a linguagem de programação	Ex: 20:20 NA: 8683
Hour	Hora em que ocorreu	Dado gerado com recurso a linguagem de programação	0-24 NA: 8683
Minutes	Hora em que ocorreu	Dado gerado com recurso a linguagem de programação	0-59 NA: 8683

Hierarquias: Decade > Year > Semester > Quarter > Month (MonthName) > Day (Weekday) > Hour > Minutes
Decade > Year > Season

Tamanho: 19221 linhas

Tabela 15. Amostra exemplificativa da tabela de dimensão Data - dimDate

DateKey int64 0 - 19220	Year object 2002 2.7% 2021 2.6% 122 others 94.8%	Month object 8 9.6% 9 9.5% 11 others 80.9%	MonthName obj... August 9.6% September 9.5% 11 others 80.9%	WeekdayName o...	Local_time object	Hour object	Minutes object
0	1900	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
1	1902	4	April	Friday	20:20	20	20
2	1902	4	April	Tuesday	Unknown	Unknown	Unknown
3	1902	10	October	Friday	Unknown	Unknown	Unknown
4	1903	4	April	Wednesday	Unknown	Unknown	Unknown
...							
19217	2023	2	February	Monday	Unknown	Unknown	Unknown
19218	2023	1	January	Thursday	Unknown	Unknown	Unknown
19219	2023	3	March	Thursday	Unknown	Unknown	Unknown
19220	2023	1	January	Friday	Unknown	Unknown	Unknown

6.2.3 Dimensão Localização - dimLocation

A dimensão localização é outra de extrema importância para o presente modelo de negócio, pois cada desastre tem de estar associado a um local (ou a vários locais). Além do referido, a importância desta dimensão torna-se muito mais elevada quando também estão associados aos locais diversas variáveis que avaliam as características desses países/regiões, como o produto interno bruto, a população, o desemprego, o turismo e a agricultura, atributos esses que são do interesse deste projeto, pois as questões produzidas futuramente na análise são em parte dependentes dos mesmos. Como referido nas restantes dimensões, foi gerada uma chave substituta (LocationKey) para identificar cada linha e se associar a um desastre. Os atributos criados nesta dimensão, provêm do dataset dos desastres naturais e dos restantes que contém as variáveis mencionadas acima. Decidiu-se não incluir a coluna “Location” como um atributo, visto que não seria útil para análises futuras considerando que os restantes atributos (PIB, população, agricultura, etc) só têm valores para o país. As colunas “RiverBasin”, “Longitude” e “Latitude” também foram retiradas pois só se aplicavam a desastres com origem na água ou a terramotos onde se eram registados os epicentros. A renomeação das colunas foi também um passo importante na criação desta dimensão, para tornar a sua interpretação mais compreensível. A coluna “GDP” foi alterada para o atributo “CountryGDP”, a coluna “Tourism” foi alterada para “CountryTourism”, a coluna “Unemployment” foi alterada para “CountryUnemployment”, a coluna “population” foi alterada para “CountryPopulation”, a coluna “Food” foi alterada para “CountryAgricultureFood” e a “Non Food” foi alterada para “CountryAgricultureNonFood”. Os restantes atributos criados foram o país (“Country”), o código do país (“CountryCode”), a região (“Region”) e o continente (“Continent”).

Após uma análise à dimensão, é possível perceber que se trata de uma dimensão de mudança lenta, visto que, devido à adição dos atributos referentes ao PIB, turismo, desemprego, etc. que são atualizados uma vez por ano, esta dimensão também só será atualizada uma vez por ano. Além disso, todas estas variáveis possuem valores anuais históricos desde 1960 o que indica que existe uma atualização por ano. Para resolver este problema, uma

estratégia do tipo 2 foi aplicada, sendo criada uma linha para cada novo ano e três novos atributos: "RowEffectiveDate", onde se detalha a data do início do ano, "RowExpirationDate", onde se detalha a data do fim do ano e "CurrentRowIndicator", onde se torna explícito se o ano é o atual ou não (através da categoria "Current" ou "Expired"). Tendo em conta que a chave substituta seria nova para cada linha gerada com este novo sistema, foi necessário atribuir uma chave supernatural para agrupar as linhas que apenas foram alterando anualmente, sendo utilizado o atributo do país, que era único nesta dimensão. Outro possível problema que esta solução criou foi a possibilidade de esta dimensão se tornar monstro, pois cada atributo foi multiplicado pelo número de anos presentes no histórico, mas a sua dimensão (13418 linhas) acabou por não passar o número de linhas da tabela de factos.

Outra filtragem que foi efetuada nesta fase foi a decisão de apenas incluir os atributos "CountryAgricultureFood" e "CountryAgricultureNonFood" do dataset da agricultura. Considerou-se que a inclusão de todas as categorias disponíveis iria tornar a dimensão bastante complexa e prejudicar a simplicidade do modelo. Foi considerado realizar um *outrigger*, no entanto, a realização de *snowflaking* não seria o indicado nesta situação, tendo em conta que esse nível de detalhe não compensaria a complexidade acrescentada ao modelo.

Os valores nulos nesta dimensão foram inevitáveis, tendo em conta que os atributos do PIB, população, desemprego, etc. só tinham dados a partir de 1960, mas os desastres estão registados desde 1900. Assim, para se manter uma coerência, mantivemos essas linhas na dimensão para que os desastres estivessem conectados sempre à linha referente ao ano a que se deu o desastre. Outros valores nulos não existem por possível falta de medição para certos países, não fazendo sentido, novamente, imputar os valores em falta. Neste caso, os valores nulos foram substituídos por "Unknown".

Por último, foi também identificada uma hierarquia nesta dimensão, começando no continente, passando pela região (parte do continente) e terminando no país que inclui todos os restantes atributos associados ao mesmo.

Tabela 16. Tabela de dimensão Localização - dimLocation. Descrição dos campos de dados: Identificadores, atributos e medidas aditivas.

Campo	Descrição dos dados	Origem dos dados	Valores
LocationKey	Chave Substituta	ID gerado sequencialmente	1-6146
Country	País em que se insere a localização	Tabela **DESASTRES NATURAIS**	Ex: Índia
CountryCode	Código do país onde ocorreu o desastre	Tabela **DESASTRES NATURAIS**	Ex: IND
Region	Região onde ocorreu o desastre	Tabela **DESASTRES NATURAIS**	Ex: Southern Asia
Continent	Continente onde ocorreu o desastre	Tabela **DESASTRES NATURAIS**	Ex: Asia
CountryGDP	Produto interno bruto, registado em dólares americanos	Tabela **GDP**	Ex: 159594492.3 NA: 3551
CountryTourism	Número de cidadãos internacionais que entram no país	Tabela **TURISMO**	Ex: 2403074088.0 NA: 9045
CountryUnemployment	Taxa de desemprego total em percentual da força de trabalho total para países no mundo.	Tabela **DESEMPREGO**	Ex: 5.543163701 NA: 7747
CountryPopulation	População total no país	Tabela **POPULATION**	Ex: 81200.0 NA: 1018
CountryAgricultureFood	Produção per capita de produtos alimentares	Tabela **AGRICULTURA**	Ex: 76.33 NA: 2875
CountryAgricultureNonFood	Produção per capita de produtos não-alimentares	Tabela **AGRICULTURA**	Ex: 163.88 NA:4131
RowEffectiveDate	Ano em que iniciou o desastre	Dado gerado com recurso a linguagem de programação	Ex: 2012-01-01
RowExpirationDate	Mês em que acabou o desastre	Dado gerado com recurso a linguagem de programação	Ex: 2012-12-31
CurrentRowIndicator	Indica o estado do desastre (se o desastre ainda está acontecer ou não)	Dado gerado com recurso a linguagem de programação	Ex: expired

Hierarquias: Continent > Region > Country (CountryCode, CountryGDP, CountryTourism, CountryPopulation, CountryAgricultureType, CountryAgricultureValue, CountryUnemployment)

Tamanho: 13418 linhas

Tabela 17. Amostra exemplificativa da tabela de dimensão Localização - dimLocation

LocationKey int64 Sorted as... 0 - 13417	Country object China 0.8% United States 0.8% 226 others 98.5%	Country_code o... CHN 0.8% USA 0.8% 226 others 98.5%	Region object Caribbean 9.1% Western Asia 8.3% 21 others 82.5%	CountryAgricult...	RowEffectiveDate	RowExpirationD...	CurrentRowIndi...
0	Cabo Verde	CPV	Western Africa	Unknown	1900-01-01	1900-12-31	expired
1	India	IND	Southern Asia	Unknown	1900-01-01	1900-12-31	expired
2	Guatemala	GTM	Central America	Unknown	1902-01-01	1902-12-31	expired
3	Canada	CAN	Northern America	Unknown	1903-01-01	1903-12-31	expired
4	Comoros	COM	Eastern Africa	Unknown	1903-01-01	1903-12-31	expired
...							
13414	Zimbabwe	ZWE	Eastern Africa	72.64	2004-01-01	2004-12-31	expired
13415	Zimbabwe	ZWE	Eastern Africa	46.68	2006-01-01	2006-12-31	expired
13416	Zimbabwe	ZWE	Eastern Africa	85.24	2012-01-01	2012-12-31	expired
13417	Zimbabwe	ZWE	Eastern Africa	113.73	2020-01-01	2020-12-31	expired

6.2.4 Dimensão Evento - dimEvent

A tabela de dimensão Evento contém informações mais detalhadas sobre cada evento, incluindo o nome atribuído, a origem, outros desastres associados, a escala e magnitude do evento, se houve ajuda humanitária e se o estado de emergência foi declarado.

Na tabela de dimensão Evento, fez-se uma renomeação das colunas para uma maior compreensão. Aqui considerámos as seguintes renomeações: 'Event_name' por 'EventName', 'Associated_dis' por 'AssociatedDisaster', 'Associated_dis2' por 'AssociatedDisaster2', 'Mag_value' por 'MagnitudeValue', 'Mag_scale' por 'MagnitudeScale' e 'OFDA_response' por 'OFDA_Response'.

A dimensão "Evento" contém 8809 linhas, com vários valores nulos. No campo "EventName", os valores nulos foram substituídos por "Name not specified", já que há desastres registados que não receberam um nome específico atribuído. Para os campos "Origin", "MagnitudeValue", "Appeal" e "Declaration", os valores nulos foram designados como desconhecidos, pois não existem registos desses dados. Para os campos "AssociatedDisaster" e "AssociatedDisaster2", os valores nulos foram substituídos por "No Disaster Associated", pois não foram registados outros desastres naturais que ocorressem simultaneamente e que fossem causados pelo desastre natural em questão. Para os campos "Magnitude", "Scale" e "OFDA_response", os valores nulos foram substituídos por "Not Applicable", pois não se aplica atribuir esses atributos em determinadas situações. Para o campo "CPI", os valores nulos foram substituídos por "Not Determined", pois em certos anos não foi calculada a variação dos preços dos bens e serviços consumidos pelas famílias ao longo do tempo do desastre em questão.

Existe uma hierarquia associada onde vai do mais abrangente, Continente, até ao mais detalhado, Country. Este tipo de hierarquia é uma hierarquia de profundidade, permitindo realizar as operações de drill-down e roll-up e escolher o nível de detalhe pretendido.

Tabela 18. Tabela de dimensão Localização - dimEvent. Descrição dos campos de dados: Identificadores, atributos e medidas aditivas.

Campo	Descrição dos dados	Origem dos dados	Valores
EventKey	Chave Substituta	ID gerado sequencialmente	1 - 8809
EventName	Nome atribuído ao evento	Tabela **DESASTRES NATURAIS**	Ex: Thomas NA: 5866
Origin	Origem ou causa do desastre	Tabela **DESASTRES NATURAIS**	Ex: Earthquake NA: 5987
AssociatedDisaster	Evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre	Tabela **DESASTRES NATURAIS**	Ex: Slide (land, mud, snow, rock) NA: 5944
AssociatedDisaster2	Outro evento relacionado que ocorre no mesmo período de tempo ou nas mesmas áreas geográficas afetadas pelo desastre	Tabela **DESASTRES NATURAIS**	Ex: Tsunami/Tidal wave NA: 8100
MagnitudeValue	magnitude ou intensidade de um desastre	Tabela **DESASTRES NATURAIS**	Ex: 145.0 NA: 4630
MagnitudeScale	informações sobre a escala utilizada para medir a magnitude	Tabela **DESASTRES NATURAIS**	Ex: Ritcher NA: 620
OFDA_Response	Intervenções da Office of U.S. Foreign Disaster Assistance (OFDA),	Tabela **DESASTRES NATURAIS**	Ex: No
Appeal	Indica se houve um apelo de ajuda humanitária	Tabela **DESASTRES NATURAIS**	Ex: No NA: 7158
Declaration	Indica se houve uma declaração oficial de estado de emergência	Tabela **DESASTRES NATURAIS**	Ex: No NA: 6524
CPI	Mede a variação dos preços de um conjunto de bens e serviços consumidos pelas famílias ao longo do tempo	Tabela **DESASTRES NATURAIS**	Ex: 3,077011162 NA: 31

Hierarquias: Origin > AssociatedDisaster > AssociatedDisaster2

Tamanho: 8809 linhas

Tabela 19. Amostra exemplificativa da tabela de dimensão Evento - dimEvent

EventKey int64 0 - 8808	EventName object Name not s... 66.6% Cholera 1.6% 1612 others 31.8%	Origin object Unknown Ori... 68% Heavy rains 8.2% 705 others 23.9%	AssociatedDisa... No Disaster... 67.5% Slide (land, ... 10.9% 30 others 21.7%	OFDA_Response o No 85.8% Yes 14.2%	Appeal object Unknown 81.3% No 16.3% Yes 2.4%	Declaration object Unknown 74.1% No 16% Yes 9.9%	CPI object 70,8487927 3.8% 66,73105799 3.5% 114 others 92.7%
0	Name not specified	Unknown Origin	Famine	No	No	No	2,849084409
1	Name not specified	Unknown Origin	No Disaster Associated	No	No	No	2,849084409
2	Name not specified	Unknown Origin	Tsunami/Tidal wave	No	Unknown	Unknown	2,963047785
3	Santa Maria	Unknown Origin	No Disaster Associated	No	Unknown	Unknown	2,963047785
4	Name not specified	Unknown Origin	No Disaster Associated	No	Unknown	Unknown	3,077011162
...	...	...	...	...	...	...	...
8805	Name not specified	Unknown Origin	No Disaster Associated	No	Unknown	Unknown	Not Determined
8806	Name not specified	Unknown Origin	No Disaster Associated	No	Unknown	Yes	Not Determined
8807	Cholera	Unknown Origin	No Disaster Associated	No	Unknown	Unknown	Not Determined
8808	Name not specified	Heavy rainfall and river overflow	No Disaster Associated	No	Unknown	Unknown	Not Determined

6.3 Diagrama em estrela

O diagrama em estrela ajuda a visualizar as relações entre as tabelas de factos e dimensões criadas e, assim, facilita a compreensão das informações armazenadas. Aqui, conseguimos visualizar a relação entre as tabelas dimLocation, dimDate, dimEvent e dimType com a tabela factDisaster.

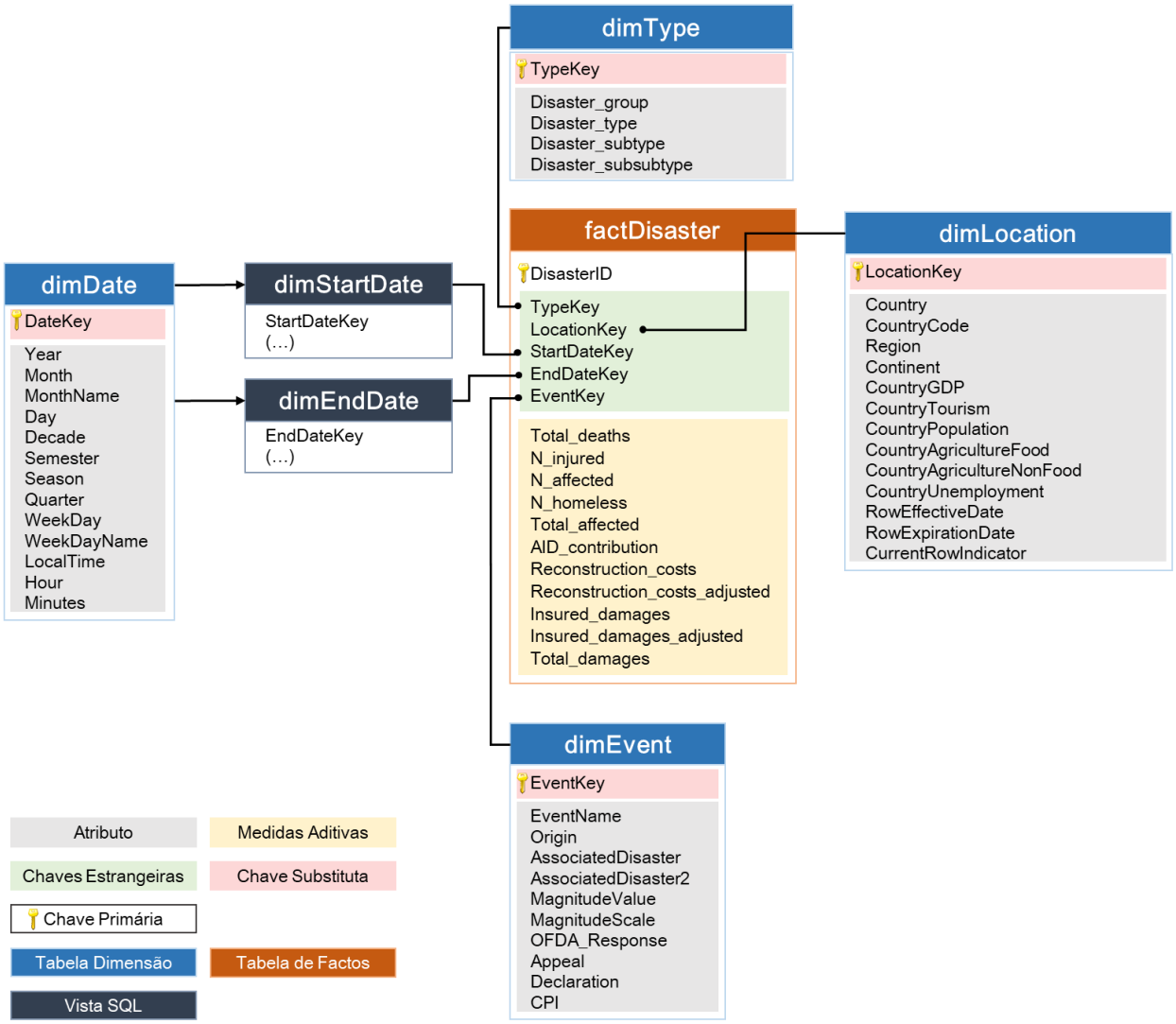

Figura 4. Diagrama em estrela para o modelo de negócio, incluindo a tabela de factos factDisaster e as respetivas dimensões.

Conclusão:

A primeira etapa incluiu a análise de fontes de dados abertas e a edição de dados existentes, como a exclusão de colunas e de observações irrelevantes para o negócio proposto. Este processo foi fundamental para aprimorar a manipulação de dados e, assim, obter informações relevantes e necessárias para a próxima etapa do projeto. Durante a primeira etapa, houve uma transformação dos dados de forma a tornar mais fácil a sua manipulação e compreensão. Além disso, nessa etapa, foi possível construir um processo de negócio e elaborar 3 questões analíticas relacionadas com o tema do projeto, cujas respostas serão elaboradas na etapa 3 do projeto.

Na segunda etapa, procedeu-se à realização da modelação dimensional para o processo de negócio criado na etapa 1 do projeto. Nesta etapa definiu-se o grão e o tipo de tabela de factos, estabeleceu-se as dimensões adequadas ao negócio, identificou-se medidas numéricas na tabela de factos e, por fim, desenhou-se o diagrama em estrela. O diagrama em estrela ajuda a visualizar as relações entre as tabelas de factos e dimensões criadas e, assim, facilita a compreensão das informações armazenadas.

Foram encontradas algumas dificuldades na construção da dimensão dimLocation, por se tratar de uma dimensão de mudança lenta, em que foi necessário ter extrema cautela na criação de novas colunas, assim como na junção de várias colunas de diversas tabelas diferentes.

Referências

1 - Negri, J. (2021). EMDAT (Emergency Events Database) - The Natural Disasters Dataset. Obtido a 18 Março de 2023, de https://www.kaggle.com/datasets/jnegrini/emdat19002021?select=EMDAT_1900-2021_NatDis.csv

2 - World Bank. (2023). GDP (current US$). Obtido a 23 Março de 2023, de https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

3 - World Bank. (2023). Population, total. Obtido a 23 Março de 2023, de https://data.worldbank.org/indicator/SP.POP.TOTL

4 - World Bank. (2023). Unemployment, total (% of total labor force). Obtido a 23 Março de 2023, de https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

5 - World Bank. (2023). International tourism, number of arrivals. Obtido a 23 Março de 2023, de https://data.worldbank.org/indicator/ST.INT.ARVL

6 - Food and Agriculture Organization of the United Nations. (2023). FAOSTAT online database. Obtido a 24 Março de 2023, de https://www.fao.org/faostat/en/#data/QI