PROJETO DE INTEGRAÇÃO E PROCESSAMENTO ANALÍTICO DE INFORMAÇÃO

Etapa 2 – Análise de Dados e Modelação Dimensional

Renato Vaz, Sílvia Mourão, Sofia Freire
fc53375,fc57541,fc53373

Conteúdo

Índice de Figuras

Índice de Tabelas

Principais alterações realizadas

Após a primeira entrega, e com atenção ao feedback recebido e às necessidades identificadas durante a segunda fase, alterámos/ acrescentámos:

- Normalizámos as análises às tabelas, realizando a análise da média, moda, mínimo, máximo e histogramas quando o tipo de dados permitia.
- Especificámos, ainda, o tipo de dados numérico presente nas tabelas.
- Acrescentámos um campo na tabela *Countries and Territories* o campo Língua do País
- Reorganizamos as tabelas *World GDP* e *World Population* para que estas apresentassem um campo *Year* e outro campo com os valores numéricos.
- De forma a enriquecer a nossa tabela de factos adicionámos um novo *dataset* ao nosso trabalho relativo à área total dos países e regiões do mundo (tabela *LandArea*).
- No diagrama relacional das fontes de dados acrescentámos os campos em falta, descritos em cima, e adicionámos a nova tabela *LandArea* (incluindo as respetivas ligações).

Introdução

O presente relatório foi elaborado no âmbito da Unidade Curricular de Integração e Processamento Analítico de Informação (IPAI).

Este trabalho irá dividir-se em três etapas. Nesta primeira etapa identificámos um tema interessante para o grupo e construímos um processo de análise de negócio através de diversos conjuntos de dados disponíveis na internet.

O tema escolhido para a realização do trabalho prático consiste numa análise sobre as tendências de voto no festival da Eurovisão ao longo dos anos. O festival da Eurovisão é um evento anual, que teve início em 1956 e teve como inspiração o Festival de Música de Sanremo. O festival da Eurovisão contará neste ano de 2022 com a sua 66º edição, tendo sido realizado ininterruptamente todos os anos desde 1956 até ao dia de hoje, exceto em 2020 devido à pandemia COVID19. O festival não teve sempre o mesmo formato, tendo existido mudanças nas regras para os participantes (como por exemplo a linguagem em que as músicas teriam de ser cantadas) e para os sistemas de votação (votação inicialmente com júri, depois por televoto e hoje em dia num sistema misto), que serão devidamente estudadas e caracterizadas numa fase posterior do trabalho.

De forma a estudar e analisar estes dados, foram recolhidos, de várias fontes de dados, diversos datasets, dividindo-se estes em duas categorias: Datasets relativos ao festival da Eurovisão e Datasets que permitem caracterizar colunas dos dados recolhidos, nomeadamente informação sobre os países, géneros de música e acontecimentos anuais.

Esta primeira fase consistiu na identificação e descrição dos conjuntos de dados bem como a análise dos valores, dos erros e das suas interligações. Por fim, com base nos dados adquiridos, definiu-se um processo de análise de negócio elaborando três questões analíticas às quais deveremos conseguir responder no final da terceira etapa.

A segunda fase do trabalho consistiu na modelação dimensional e preparação dos dados para integrarem um data warehouse que será construído na terceira fase. Identificado o processo de negócio durante a primeira fase, decidimos que seria necessário criar duas tabelas de facto de forma a dar resposta às questões analíticas que tinham sido colocadas aquando da descrição deste processo. Foi identificado o grão destas tabelas, assim como as tabelas de dimensão que fazem sentido incluir em cada uma delas e aquelas que poderão ser partilhadas entre as duas. Finalmente, foram identificadas e descritas as medidas numéricas que irão fazer parte da tabela de factos e desenhou-se o esquema em estrela que servirá de base ao sistema ETL que será implementado na terceira etapa do trabalho.

1. Fontes de dados

A primeira fase do projeto consistiu na pesquisa e recolha de dados relevantes para o tema escolhido, assim como dados que podem ser utilizados para os caracterizar. Os dados utilizados para a elaboração deste projeto são provenientes de diversas fontes, descritas na tabela seguinte.

Tabela 1- Conjunto de dados e a sua descrição

Dataset	Descrição	Formato	Extração	Link
Eurovision finals voting results 1957-2021 "Eurovisao.xlsx"	Dataset retirado do Kaggle com os votos de todos os países na Eurovisão de 1957 a 2021	.csv	Download direto	https://www.kaggle.com/orianao/eurovision-finals-voting-results-19572021
Top Spotify (2010-2019) "Genero.xlsx"	Dataset retirado do Kaggle com as músicas mais ouvidas por ano (2010-2019) no mundo pelo Spotify, com dados baseados na Billboard	.csv	Download direto	https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year
Top Spotify (2020-2021) "Genero.xlsx"	Dataset retirado do Kaggle com as músicas mais ouvidas do Spotify nos anos 2020 e 2021.	.csv	Download direto	https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021
Land Borders (Neighbours) "Vizinhos.xlsx"	Lista de países e territórios por fronteiras terrestres. Inclui o número de fronteiras terrestres distintas de cada país ou território, bem como os nomes de seus países e territórios vizinhos.	.xlsx	Tabela de página convertida em tabela Excel	https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders
Eurovision song lyrics 1956-2021 "Musica.xlsx"	Dataset retirado do Kaggle com dados sobre as músicas que participaram nos festivais da Eurovisão entre 1956 e 2021	. json	Download direto, importado e convertido em tabela Excel	https://www.kaggle.com/minitree/eurovision-song-lyrics
Countries and Territories "Localizacao.xlsx"	Dataset que exibe os limites administrativos mundiais de nível 0. Contém uma lista dos países, bem como territórios não soberanos	.xlsx	Download direto	https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/table/
Tourist "Turistas.xlsx"	Dataset retirado do Eurostat que exibe resultados do número de chegadas de turistas a estabelecimentos de alojamento turístico por país (com base em hotéis e acomodações similares)	.xls	Download direto	https://ec.europa.eu/eurostat/databrowser/view/tour_occ_arnat/default/table?lang=en

Historical emissions "Emissoes.xlsx"	Dataset retirado do Kaggle com emissões históricas de dióxido de carbono ao longo de 3 décadas, por todos os países do mundo	.csv	Download direto	https://www.kaggle.com/datasets/ankanhore545/carbon-dioxide-emissions-of-the-world
World GDP (PIB 1960-2020) "PIB.xlsx"	Dataset retirado do Kaggle com os valores do PIB, o crescimento do PIB, o PIB per capita e o crescimento do PIB per capita para todos os países do mundo entre 1960 e 2020.	.csv	Download direto	https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths
World Population (1960-2020) "Populacao.xlsx"	Dataset retirado do *The World Bank* com informação sobre a população de todos os países e regiões do mundo, de 1960 até 2020.	.xls	Download direto	https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=WLD#
Conflicts Participants "Conflitos.xlsx"	Dataset retirado do Kaggle com uma lista de conflitos mundiais ocorridos depois do final da Segunda Guerra Mundial e os países envolvidos nesses conflitos	.csv	Download direto	https://www.kaggle.com/datasets/guybarash/war-conflicts-and-nations-who-took-part-in-them?select=Conflicts+participants.csv
Land Area "Area.xlsx"	Dataset retirado do *The World Bank* com informação sobre as áreas de todos os países e regiões do mundo, de 1961 até 2021	.xls	Download direto	https://data.worldbank.org/indicator/AG.LND.TOTL.K2

2. Descrição e análise dos datasets

A fase seguinte do trabalho consistiu em descrever e analisar os dados recolhidos, incluindo uma validação dos valores contidos nos datasets e posterior correção de erros encontrados. Nos casos onde existiam lacunas nos dados foram ainda atualizadas as tabelas ou criados campos para permitir uma melhor integração entre tabelas no futuro.

2.1 Eurovisão

O ficheiro "*Eurovisão.xlsx*" contém dados e informação relevante sobre a Eurovisão. Nesta tabela encontram-se dados sobre a cidade onde foi realizada a edição da Eurovisão, o ano, o tipo de pontos, o país que deu pontos, o país que recebeu pontos e a quantidade de pontos recebidos, descritos na seguinte tabela.

Tabela 2 - Descrição dos campos do dataset da Eurovisão

#	Campo	Tipo de dados	Descrição	Exemplo
1	Index	Categórico	Identificador único	544
2	Host City	Texto	Cidade anfitriã da Eurovisão	Tel Aviv
3	Year	Número	Ano em que se realizou a Eurovisão	2019
4	Points type	Categórico	Tipo de pontos dados	Points given by televoters
5	From	Texto	País que deu os pontos	Albania
6	To	Texto	País que recebeu os pontos	Italy
7	Points	Número	Número de pontos que um certo país deu a outro	8

2.1.1 Análise estatística

Para esta tabela considerou-se apenas a análise estatística das máximas e mínimas dos campos "*Index*", "*Year*" e "*Points*". No que toca aos campos de texto, apenas fez sentido investigar qual a moda dos resultados e qual o número de ocorrências para essas modas.

2.1.1.1 – Index

Tabela 3 - Descrição do campo "Index" do dataset da Eurovisão

Campo	Máximo	Mínimo
Index	13446	0

O campo "*Index*" tem o máximo de 13446 que corresponde ao número total de instâncias de um país a dar pontos a outro país.

2.1.1.2 – Host City

Tabela 4 - Descrição do campo "Host City" do dataset da Eurovisão

Campo	Moda	Ocorrências
Host City	Dublin	6

O campo “Host City” designa a cidade anfitriã da Eurovisão. Através desta análise foi possível concluir que a cidade que mais vezes acolheu o festival foi Dublin, um total de 6 vezes.

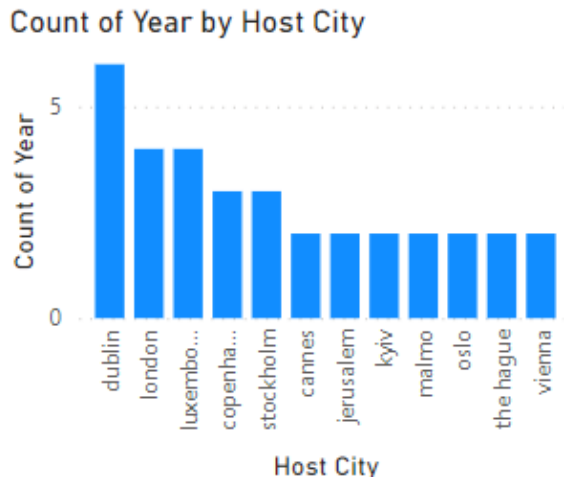

Figura 1 - Moda de cidade anfitriã

2.1.1.3 – Year

Tabela 5 - Descrição do campo "Year" do dataset da Eurovisão

Campo	Máximo	Mínimo
Year	2021	1957

O campo “*Year*” mostra a extensão temporal dos dados, que vão desde 1957 até 2021.

2.1.1.4 – Points Type

Tabela 6 - Descrição do campo "Points Type" do dataset da Eurovisão

Campo	Valores Possíveis	Ocorrências
Points Type	“Points given by televoters”	1290
	“Points given by the jury”	1290
	“Points given”	10867

O campo tipo de pontos distingue pontos dados por júri e pontos dados por televoto, existindo um terceiro campo para “Points given” que corresponde aos anos antes de existir uma distinção entre os tipos de pontos.

2.1.1.5 – From

Tabela 7 - Descrição do campo "From" do dataset da Eurovisão

Campo	Moda	Ocorrências
From	United Kingdom	624

O campo “From” designa o país que dá pontos. O país que mais vezes deu pontos foi o Reino Unido. É de notar que o top quatro na lista do “*From*” são os países que tem qualificação direta para a final e que têm participado de forma regular no concurso, Reino Unido, Espanha, Alemanha e Franca, estando em falta nesta lista a Itália que não participou durante um longo período nos anos 90.

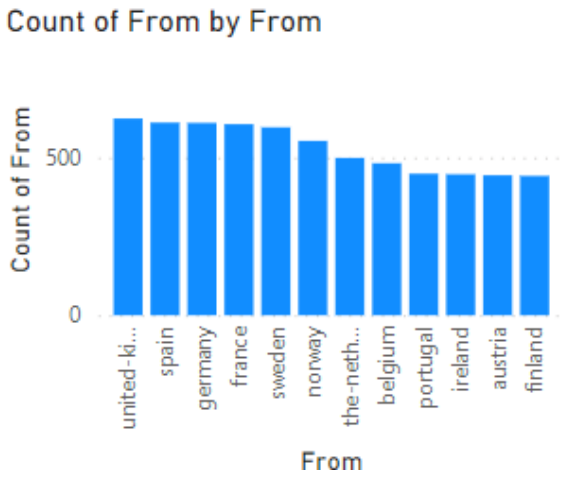

Figura 2 - Moda de votos dados

2.1.1.6 – To

Tabela 8 - Descrição do campo "To" do dataset da Eurovisão

Campo	Moda	Ocorrências
To	Sweden	721

O campo "To" refere-se ao país que recebe os pontos. Desta tabela é possível verificar que o país que mais vezes recebeu pontos foi a Suécia. A Suécia é um dos países com o segundo maior número de vitórias, sendo que muitas destas ocorreram recentemente, quando existia um maior número de países a dar pontos.

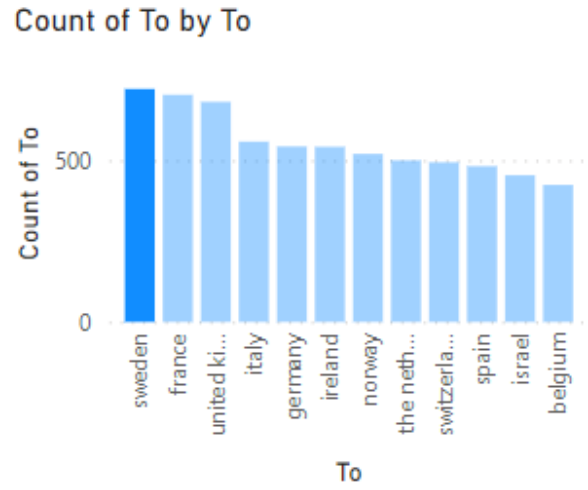

Figura 3 - Moda de votos recebidos

2.1.1.7 - Points

Tabela 9 - Descrição do campo "Points" do dataset da Eurovisão

Campo	Máximo	Mínimo
Points	12	1

O campo "*Points*" mostra o intervalo dos pontos possíveis, entre 1 e 12. Neste caso, a tabela mostra apenas as instâncias onde um país deu pontos a outro país, mas não mostra as combinações de países que não deram pontos uns aos outros.

2.1.2 Erros e dados em falta

Não foram detetados erros nesta tabela. Não existem dados sobre a votação da primeira edição do festival da Eurovisão.

2.2 Top Spotify 2010-2019

O ficheiro "*TopSpotify2010-2019.csv*" contém informação sobre as tendências musicais entre 2010 e 2019. Após realizar download do ficheiro, este inicialmente continha os seguintes campos: [*ID, Title, Artist, Top genre, Year, BPM, nrgy, dnce, dB, live, Val, dur, acous, spch, pop*].

Contudo, dado que o objeto de estudo deste dataset era analisar os géneros musicais mais populares ao longo do ano, eliminámos os campos que considerámos irrelevantes, ficando com os seguintes, estando estes representados na tabela abaixo. Estes foram depois adicionados à folha Excel "*Generos.xlsx*".

Tabela 10 - Descrição dos campos do dataset Spotify 2010-2019

#	Campo	Tipo de dados	Descrição	Exemplo
1	Index	Categórico	Identificador único	1
2	Title	Texto	Nome da música	"Hey, Soul Sister"
3	Artist	Texto	Nome do artista da música	Train
4	Top genre	Texto	O género da música	Dance Pop
5	Year	Número	Ano da música no Billboard	2017
6	Pop	Número	Popularidade	83

2.2.1 Análise estatística

Para esta tabela considerou-se apenas a análise estatística das máximas e mínimas dos campos "*Index*", "*Year*" e "pop". No que toca aos campos de texto, apenas fez sentido investigar qual a moda dos resultados e qual o número de ocorrências para essas modas. Foi ainda feita uma análise de qual o género mais popular por ano. Devido a alguns erros nos dados, descritos abaixo, para que cada ano tivesse o mesmo número de músicas mais ouvidas, reduzimos as amostras a 31

2.2.1.1 – Index

Tabela 11 - Descrição do campo "Index" do dataset Spotify 2010-2019

Campo	Máximo	Mínimo
Index	603	1

O campo "*Index*" tem o máximo de 603 que corresponde ao número total de instâncias de músicas na tabela. No entanto, foram apenas consideradas para a análise final 310 músicas.

2.2.1.2 – Title

O campo "*Title*" é um campo de texto único, sem repetições, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.2.1.3 – Artist

Tabela 12 - Descrição do campo "Artist" do dataset Spotify 2010-2019

Campo	Moda	Ocorrências
Artist	"Bruno Mars"	11

	“Katy Perry” “Maroon 5”	11 11

A moda do campo “*Artist*” representa os artistas que mais vezes aparecem no dataset, ou seja, os artistas com maior número de músicas no top de popularidade.

2.2.1.4 – Top Genre

Tabela 13 - Descrição do campo "Top Genre" do dataset Spotify 2010-2019

Campo	Moda	Ocorrências
Top Genre	“dance pop”	149

O campo “*Top Genre*” refere-se ao género predominante da música popular. A moda do dataset, ou seja, o género mais popular, é dance pop, com 149 ocorrências.

2.2.1.5 – Year

Tabela 14 - Descrição do campo "Year" do dataset Spotify 2010-2019

Campo	Máximo	Mínimo
Year	2019	2010

O campo “*Year*” mostra a extensão temporal dos dados, que vão desde 2010 até 2019.

2.2.1.6 – pop

Tabela 15 - Descrição do campo "pop" do dataset Spotify 2010-2019

Campo	Máximo	Mínimo	Moda	Média
pop	99	59	78	76.06

O campo “*pop*” é um indicador de popularidade, que varia entre 59 e 99. A moda deste campo é 78 e a média é 76.06.

2.2.1.7 – Género mais popular por ano

Realizámos a moda para cada ano, com o objetivo de perceber qual o género musical mais ouvido em cada ano.

Tabela 16 - Moda e número de ocorrências para o género musical mais ouvido entre os anos 2010 e 2019.

Ano	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Moda	Dance pop	Dance pop	Dance pop	Dance pop	Dance pop	Dance pop	Dance pop	Dance pop	Dance pop	Pop
Ocorrências	22	21	13	15	14	13	17	12	16	9

Analisando a tabela, conseguimos concluir que o estilo musical mais ouvido no período temporal entre 2010 e 2019 é o Dance Pop à exceção do ano 2019, onde o estilo mais ouvido é o Pop.

2.2.2 Erros e dados em falta

Ao observarmos os dados adquiridos, reparámos que havia erros pois existiam músicas com popularidade igual a zero. Para que esses dados não contaminassem as análises futuras, decidimos eliminar esses valores da tabela.

2.3 Top Spotify 2020-2021

Dado que o dataset anterior não continha dados relativos ao ano 2021, sentimos a necessidade de encontrar dados relativos ao ano em causa. O ficheiro "*TopSpotify2020-2021.csv*" continha, inicialmente os seguintes campos: *[Index, Highest Charting Position, Charting Position, Number of Times Charted, Week of Highest Charting, Song Name, Streams, Artist, Artist Followers, Song ID, Genre, Release Date, Weeks Charted, Popularity, Danceability, Energy, Loudness, Speechiness, Acousticness, Liveness, Tempo, Duration (ms), Valence, Chord].* Para que as duas tabelas contenham a mesma informação, eliminámos os campos que não eram comuns e que eram irrelevantes para o objeto de estudo, ficando com os restantes, visíveis na tabela seguinte. Foram ainda eliminados os dados relativos a 2020, que não iriam ser utilizados no projeto. Estes valores foram depois combinados com os valores do dataset anterior na tabela "*género.xlsx*".

Tabela 17 - Descrição dos campos do dataset Spotify 2020-2021

#	Campo	Tipo de dados	Descrição	Exemplo
1	Index	Categórico	Identificador único	1
2	Song Name	Texto	Nome da música	Beggin'
3	Artist	Texto	Nome do artista da música	Måneskin
4	Genre	Lista	Género da música	Indie rock
5	Weeks Charted	Data	Semanas em que a música teve no Top do Spotify	2021-07-23--2021-07-30
6	Popularity	Número	Popularidade da música	100

2.3.1 Análise estatística

Para esta tabela considerou-se apenas a análise estatística das máximas e mínimas dos campos "*Index*" e "*Popularity*". No que toca aos campos de texto, apenas fez sentido investigar qual a moda dos resultados e qual o número de ocorrências para essas modas. Foi ainda feita uma análise de qual o género mais popular por ano. À semelhança dos dados anteriores, reduzimos as amostras a 31 e realizámos a moda apenas para o ano 2021.

2.3.1.1 – Index

Tabela 18 - Descrição do campo "Index" do dataset Spotify 2020-2021

Campo	Máximo	Mínimo
Index	31	1

O campo "*Index*" tem o máximo de 31 que corresponde ao número total de instâncias de músicas na tabela, após redução dos intervalos de tempo e das amostras.

2.3.1.2 – Song Name

O campo "*Song Name*" é um campo de texto único, sem repetições, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.3.1.3 – Artist

Tabela 19 - Descrição do campo "Artist" do dataset Spotify 2020-2021

Campo	Moda	Ocorrências
Artist	"Olivia Rodrigo"	5

A moda do campo "*Artist*" representa os artistas que mais vezes aparecem no dataset, ou seja, o artista com maior número de músicas no top de popularidade.

2.3.1.4 – Genre

Tabela 20 -Moda e número de ocorrências para o género musical mais ouvido em 2021

Campo	Moda	Ocorrências
Genre	Pop e Hip Hop	6

Ao observarmos a tabela acima, observamos que houve dois estilos musicais que mais se repetiram no número de músicas mais ouvidas no Spotify, sendo esses o Pop e o Hip Hop, cada um com 6 ocorrências.

2.3.1.5 – Weeks Charted

O campo "*Weeks Charted*" é um campo que descreve quais as semanas em que estas músicas estiveram no top, o que corresponde ao intervalo entre 3/1/2021 a 26/12/2021.

2.3.1.6 – Popularity

Tabela 21 - Descrição do campo "Popularity" do dataset Spotify 2020-2021

Campo	Máximo	Mínimo	Moda	Média
Popularity	100	85	92	93.48

O campo "Popularity" é um indicador de popularidade, que para os valores considerados varia entre 85 e 100. A moda deste campo é 92 e a média é 93.48.

2.1.2 Erros e dados em falta

Não foram encontrados erros nem dados em falta neste dataset.

2.4 Vizinhos

Para se obter os dados dos vizinhos de cada país foi criada uma tabela no Excel através de dados recolhidos com recurso a uma página da Wikipédia (Wikipedia, 2022) e ao Google Maps, uma vez que estes sites não permitem exportar de forma direta os dados num formato .csv ou Excel. Com os dados já num formato Excel foi possível eliminar alguns campos que não iremos analisar, como por exemplo o comprimento da fronteira entre países (em Km e milhas). Os campos considerados encontram-se descritos na tabela seguinte.

Tabela 22 - Descrição dos campos do dataset Vizinhos

#	Campo	Tipo de dados	Descrição	Exemplo
1	ID	Categórico	Identificador único	1
2	Country	Texto	Nome do País	Portugal
3	No. Of Neighbours	Número	Número de vizinhos	1
4	Neighbours	Texto	Lista dos países vizinhos	Spain

2.4.1 Análise estatística

A tabela "Vizinhos.xlsx" foi editada de forma a conter apenas informação sobre os países da Eurovisão, que nos iriam interessar para a próxima fase do trabalho.

2.4.1.1 – ID

Tabela 23 - Descrição do campo "ID" do dataset Vizinhos

Campo	Máximo	Mínimo
ID	49	1

O campo "*ID*" tem o máximo de 49 que corresponde ao número total de instâncias de países existentes na tabela.

2.4.1.2 – Country

O campo "*Country*" é um campo de texto único, sem repetições, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.4.1.3 – No. Of Neighbours

Tabela 24 - Descrição do campo "No. of Neighbours" do dataset Vizinhos

Campo	Média	Moda	Mínimo	Máximo
No. of Neighbours	3.53	4	0	10

O número de vizinhos na Europa varia entre 0-10, com uma média de 3.53 vizinhos por país. Para se ter uma ideia da contagem de países que partilham o mesmo número de países vizinhos fizemos um histograma do campo "*No. Of Neighbours*".

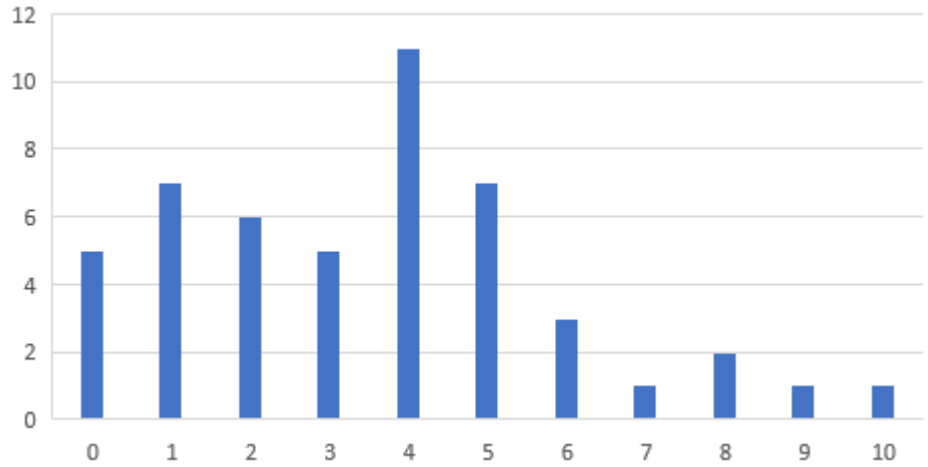

Figura 4 - Histograma do número de vizinhos

2.4.1.3 - Neighbours

O campo "Neighbours" continha uma lista de todos os vizinhos de um país numa só célula. Poderia ter sido feita a análise da moda separando esta lista em colunas diferentes, no entanto o país que corresponderia à moda seria o país com o maior número de vizinhos, o que pode ser determinado apenas por uma análise simples da tabela.

Tabela 25 - Descrição do campo "Neighbours" do dataset Vizinhos

Campo	Moda	Ocorrências
Neighbours	Rússia	10

2.4.2 Erros e dados em falta

Esta tabela foi construída para o projeto e por isso não existem erros ou dados em falta.

2.5 Música

O ficheiro *"Musica.xlsx"* contém dados e informação sobre as músicas que participaram no festival da Eurovisão desde 1956, incluindo dados sobre o país, o artista e a língua em que a música é cantada. Desta tabela foram eliminados os campos correspondentes à letra da música original e letra da música traduzida pois não são relevantes para o projeto.

Tabela 26 - Descrição dos campos do dataset Música

#	Campo	Tipo de dados	Descrição	Exemplo
1	ID	Categórico	Identificador único	176
2	Country	Texto	Nome do País	Portugal
3	Participation	Categórico	Número cumulativo de participações do país	5
4	Artist	Texto	Nome do artista	Carlos Mendes
5	Song	Texto	Nome da música	Verão
6	Language	Texto	Linguagem da música	Portuguese
7	Pl.	Categórico	Classificação Final	11
8	Sc.	Número	Pontuação final	5
9	Eurovision_Number	Categórico	Número de edição	13
10	Year	Número	Ano da edição	1968
11	Host_Country	Texto	Pais organizador	United Kingdom
11	Host_City	Texto	Cidade onde foi realizado o festival	London

Dado que a tabela anterior não apresentava dados suficientes, sentimos a necessidade de a completar acrescentando as seguintes colunas. A forma como estas foram geradas é descrita na secção dos erros e dados em falta.

Tabela 27 - Descrição dos campos adicionados ao dataset Música

#	Campo	Tipo de dados	Descrição	Exemplo
12	EnglishNonEnglish	Categórico	Indica se a linguagem da música é inglês, não inglês ou mistura.	English
13	RunningOrder	Número	Número na ordem do concurso em que a música tocou.	1
13	ROpercent	Categórico	Razão entre a ordem dentro do concurso em que a música tocou e o número total de músicas na final.	0.64
15	Genre	Texto	Género musical	Dance Pop

2.5.1 Análise estatística

Na análise deste dataset foram consideradas modas para os dados no formato texto e algumas medidas estatísticas onde relevantes para os dados numéricos.

2.5.1.1 – ID

O campo "*ID*" tem o máximo de 1644 que corresponde ao número total de músicas que participaram no festival da Eurovisão.

Tabela 28 - Descrição do campo "ID" do dataset Música

Campo	Máximo	Mínimo
ID	1644	1

2.5.1.2 – Country

Na tabela seguinte é possível verificar que o país que mais vezes participou na Eurovisão foi a Alemanha, com um valor total de participações de 64, tendo apenas ficado de fora um ano.

Tabela 29 - Descrição do campo "Country" do dataset Música

Campo	Moda	#
Country	Germany	64

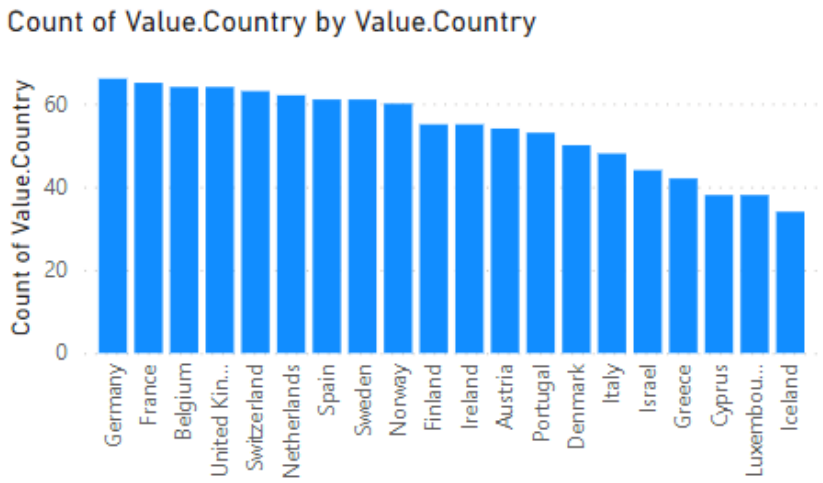

Figura 5 - Países com mais participações na Eurovisão a partir do dataset música da Eurovisão

2.5.1.3 – Participation

Tabela 30 - Descrição do campo "Participation" do dataset Música

Campo	Máximo	Mínimo
Participation	64	1

O máximo de participações por um país na Eurovisão é 64, obtido pela Alemanha como foi visto acima. O mínimo de participações é 1, no caso de Marrocos que participou apenas uma vez.

2.5.1.4 – Artist

Pela análise dos dados podemos também procurar quais os artistas que mais vezes representaram o seu país na eurovisão, existindo quatro artistas com quatro participações cada.

Tabela 31 - Descrição do campo "Artist" do dataset Música

Campo	Moda	Ocorrências	País Representado
Artist	Fud Leclerc	4	Belgium
	Lys Alyssa	4	Switzerland
	Peter, Sue and Marc	4	Switzerland
	Valentina Monetta	4	San Marino

Uma nota sobre a tabela acima, Lys Alyssa cantou por quatro vezes na Eurovisão, no entanto duas destas participações aconteceram no primeiro ano do festival, onde excecionalmente cada país apresentou duas músicas.

2.5.1.5 – Song

O campo "*Song*" é um campo de texto único. Embora possam existir repetições de títulos estes não correspondem à mesma música, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.5.1.6 – Language

No que toca à análise das linguagens utilizadas nas músicas a concurso, foram analisados os campos *"Value.Language"*, que contém a linguagem original, e o campo *"EnglishNonEnglish"* que apenas verifica se a linguagem é inglês, não inglês ou uma mistura de línguas. Neste caso, a linguagem individual mais popular na Eurovisão é o inglês.

Tabela 32 - Descrição do campo "Language" do dataset Música

Campo	Moda	Ocorrências
Language	English	615

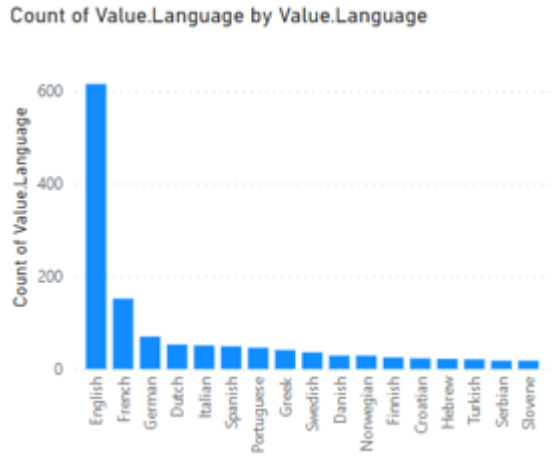

Figura 6 - Campo de linguagem do dataset música da Eurovisão

2.5.1.7 – Pl.

O campo *"pl"* representa o lugar em que uma música terminou o concurso. A sua média não corresponde ao valor central entre o máximo e o mínimo devido à variação do número de participantes ao longo dos anos.

Tabela 33 - Descrição do campo "Pl" do dataset Música

Campo	Média	Máximo	Mínimo	Desvio Padrão
Pl	11.26	26	1	6.68

2.5.1.8 – Sc.

Para o caso do resultado final, foram analisados os valores numéricos dos pontos recebidos por cada música, sendo necessário no entanto referir que a quantidade de pontos dada por cada país tem vindo a variar ao longo do tempo, começando por uma atribuição de pontos de 1 a 5 na primeira edição até ao atual sistema de voto em vigor desde 2016 em que cada país apresenta dois conjuntos de pontos para dar, um por decisão de um júri interno e outro por televoto, sendo estes conjuntos compostos pelos seguintes números de pontos: [1,2,3,4,6,7,8,10,12].

O valor máximo de pontos foi atingido em 2017 com a música "Amar pelos Dois", interpretada por Salvador Sobral por Portugal. No lado oposto do espectro são várias as músicas e os países que obtiveram 0 pontos no concurso. A pontuação media é de 71 pontos, existindo um elevado desvio padrão.

Tabela 34 - Descrição do campo "Sc" do dataset Música

Campo	Média	Máximo	Mínimo	Desvio Padrão
Sc	71	758	0	81

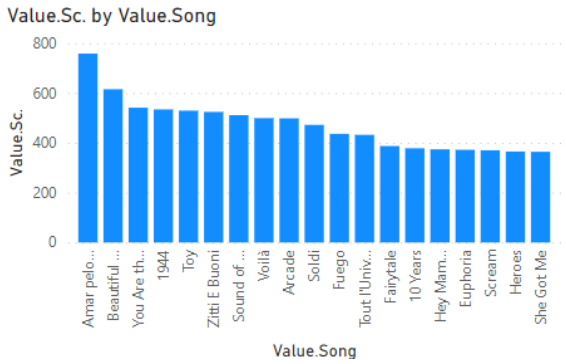

Figura 7 - Músicas com pontuações mais altas no dataset música da Eurovisão

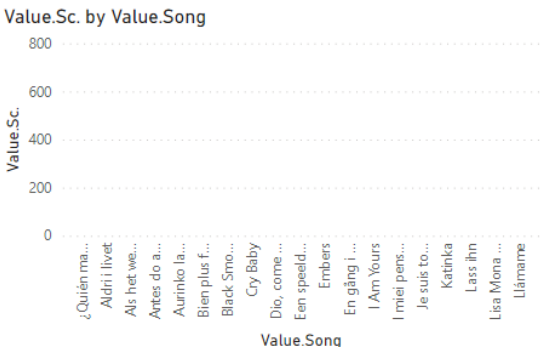

Figura 8 - Músicas com pontuações mais baixas no dataset música da Eurovisão

2.5.1.9 - Eurovision_Number

A tabela contém dados que variam entre as edições número 1 e número 65 da Eurovisão. Existe um possível problema nos dados, que seria o facto do número 65 ter sido utilizado para a edição do ano 2020, que foi depois cancelada, e para a edição do ano 2021, no entanto os dados de 2020 não vão ser analisados no âmbito deste projeto.

Tabela 35 - Descrição do campo "Eurovision_Number" do dataset Música

Campo	Máximo	Mínimo
Eurovision_Number	65	1

2.5.1.10 – Year

O intervalo temporal dos dados varia entre 1956 e 2021.

Tabela 36 - Descrição do campo "Year" do dataset Música

Campo	Máximo	Mínimo
Year	2021	1956

2.5.1.11 - Host_Country e Host City

Esta análise corresponde de dados recaiu sobre *"Host Country"* e *"Host City"*, que já tinha anteriormente sido determinada com utilização do outro dataset sobre a Eurovisão. Neste caso, confirma-se o facto anteriormente discutido sobre Dublin anfitriã do maior número de edições do festival, enquanto, à escala do país, o Reino Unido conta com o maior número de ocorrências.

Tabela 37 - Descrição dos campos "Host Country" e "Host City" do dataset Música

Campo	Moda	Ocorrências
Host Country	United Kingdom	8
Host City	Dublin	6

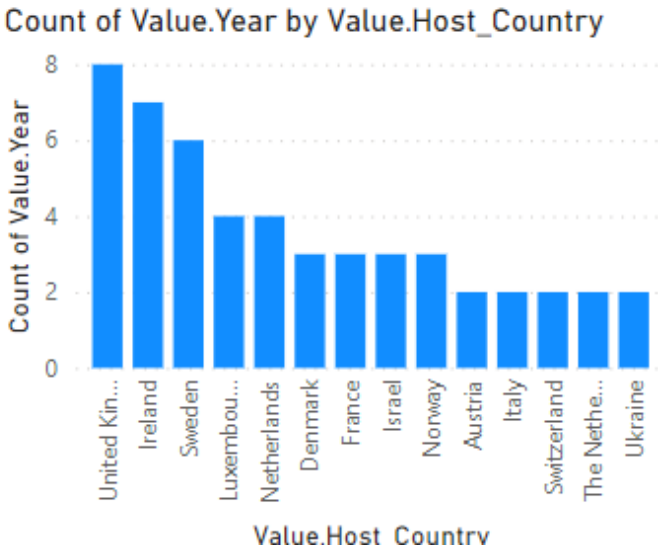

Figura 9 - Países anfitriões do maior número de edições no dataset música da Eurovisão

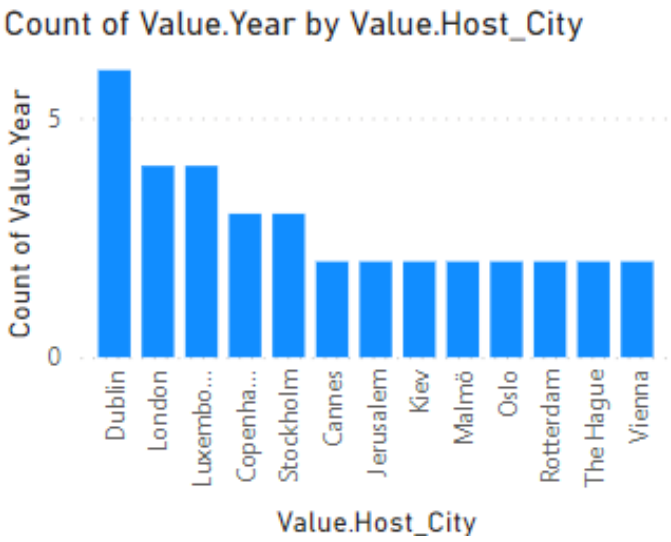

Figura 10 - Cidades anfitriãs do maior número de edições no dataset música da Eurovisão

2.5.1.12 – EnglishNonEnglish

Após analisar a linguagem individual da música, foi também analisado o campo *"EnglishNonEnglish"* que apenas verifica se a linguagem é inglês, não inglês ou uma mistura de línguas. Embora, como visto anteriormente, a linguagem individual mais popular na Eurovisão seja o inglês, existe um maior número de ocorrências somadas de outras linguagens.

Tabela 38 - Descrição do campo "EnglishNonEnglish" do dataset Música

Campo	Moda	Ocorrências
EnglishNonEnglish	Not English	844

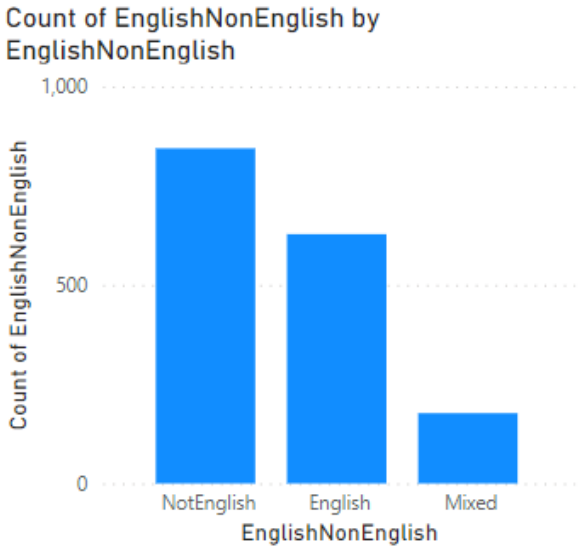

Figura 11 - Campo de validação da linguagem do dataset música da Eurovisão

2.5.1.13 – RunningOrder e ROpercent

Da analise anterior do campo "*pl*", e do campo *"running order"* é possível visualizar uma particularidade nestes dados. Para estes dois campos seria de esperar o mesmo valor de máximo e mínimo, visto que só podem existir tantos classificados quantos o número de atuações, no entanto isto não se trata de um erro, mas sim de um caso de *outlier*. A edição de 2015 celebrou os 60 anos do festival da Eurovisão e contou por isso com 1 país convidado – Austrália, o que elevou o número total de participantes para 27. Ainda assim, o valor máximo da classificação é 26 pois nesse ano dois países terminaram empatados em último lugar.

Tabela 39 - Descrição do campo "Running Order" do dataset Música

Campo	Media	Máximo	Mínimo	Desvio Padrão
Running Order	11.34	27	1	6.67

Já o campo *"ROpercent"* foi criado a partir da divisão do campo "*Running Order*" pelo total de participantes numa edição, pelo que os seus valores vão variar entre 0 e 1 e a sua analise estatística não terá significado.

2.5.1.14 – Genre

No que toca ao género musical estes dados foram apenas analisados desde o ano 2010, de forma a corresponderem com os outros datasets relacionados com o género musical e descritos anteriormente. O estilo musical mais predominante na Eurovisão é Pop.

Tabela 40 - Descrição do campo "Genre" do dataset Música

Campo	Moda	# Occurences
Genre	Pop	122

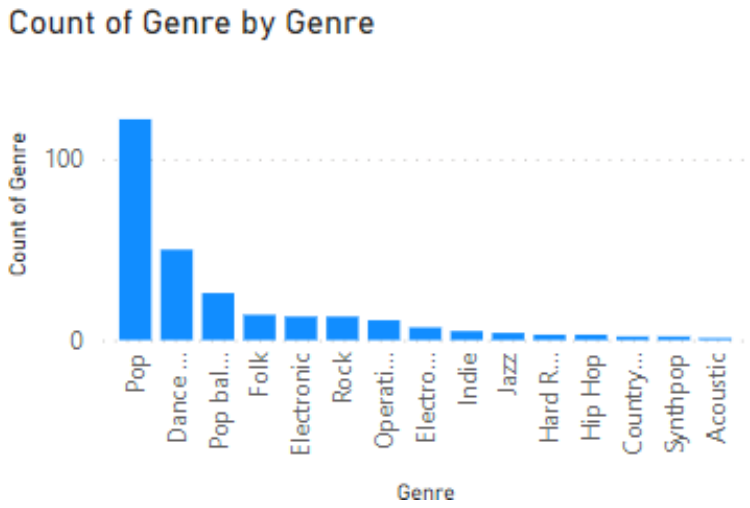

Figura 12 - Distribuição do género musical do dataset música da Eurovisão

2.5.2 Erros e dados em falta

Relativamente aos dados em falta, estes foram adicionados do seguinte modo:

O Campo *"EnglishNonEnglish"* foi gerado através de uma função Excel de "*If clause*" para a coluna linguagem. Os valores "*Mixed"* - mais do que uma linguagem utilizada na música - foram adicionados posteriormente, assim como algumas correções ao campo, como, por exemplo, músicas em inglês com títulos numa outra linguagem que não foram assumidas como inglês pelo programa.

O campo "*ContestRunningOrder*" foi acrescentado manualmente a partir de dados encontrados no site da Eurovisão (Eurovision, 2022).

O campo "*Genres*" foi acrescentado a partir de dados recolhidos com recurso a uma base de dados de música online (Discogs, 2022).

O campo "*ROpercent*" foi criado para normalizar a ordem das músicas com vista a facilitar a posterior análise dos dados. O problema está no facto de diferentes edições da Eurovisão terem sido realizadas com diferentes números de participantes, pelo que se pretendermos fazer uma análise como por exemplo "será que as músicas que tocam na segunda metade do concurso têm uma melhor classificação?" o conceito de metade depende do número total de participantes. Como tal, foi feita uma divisão da ordem das músicas no concurso pelo número total de músicas.

Relativamente aos erros e possíveis problemas que podemos encontrar no futuro:

No que toca à organização do concurso, o primeiro problema encontrado foi que, como não existiam critérios de desempate nas edições mais antigas da eurovisão gerou-se o problema de existir lugares repetidos na ordem final dos participantes. Em 1969 existiram 4 vencedores.

Alem disso, a Eurovisão de 1956 (primeira edição) é um *outlier* em relação aos outros concursos, pois foi realizado num formato um pouco diferente. Cada país a concurso apresentou duas músicas e não existem dados sobre "*place*" e "*score*" pois estes não foram revelados, apenas o vencedor.

Uma outra questão é o facto de a Eurovisão de 2020, que não se realizou devido à pandemia COVID19. Os dados sobre as músicas e artistas constam da tabela, mas obviamente não têm resultados do concurso.

Existiam ainda alguns pequenos erros na informação da tabela, como o facto de esloveno aparecer como "slovene" e "slovenian" e o facto dos Paises Baixos aparecerem como "Netherlands" e "The Netherlands". Para o resolver, substituímos os valores de "slovenian" para "slovene" e convertemos os campos "The Netherlands" em "Netherlands".

2.6 Localização

A tabela *"Localizacao.xlsx"*, tal como importada, incluía alguns campos que não eram pertinentes para a análise em estudo por isso procedemos à eliminação dos mesmos. Em baixo apresentamos os dados e informações mais relevantes desta tabela.

Tabela 41 - Descrição dos campos do dataset Localização

#	Campo	Tipo de dados	Descrição	Exemplo
1	ID	Categórico	Identificador único do país	37
2	Country	Texto	Nome do País	Portugal
4	Continent	Texto	Designação continente a que pertence o país.	Europe
5	Region	Texto	Designação da região do país dentro do continente.	Southern Europe
	EnglishLanguage	Categórico	Indica se o país tem como linguagem oficial inglês ou não.	"English" "NotEnglish" "Mixed"

2.6.1. Análise estatística

De modo a analisar esta tabela foram determinadas as modas dos campos de texto, visto que não existem campos numéricos que permitam uma análise mais detalhada. Os dados existentes na tabela foram já filtrados de forma a conter apenas países que já participaram na Eurovisão.

2.6.1.1 - ID

Tabela 42 - Descrição do campo "ID" do dataset Localização

Campo	Máximo	Mínimo
ID	49	1

O campo "*ID*" tem o máximo de 49 que corresponde ao número total de países que já participaram na Eurovisão.

2.6.1.2 – Country

O campo "*Country*" é um campo de texto único, sem repetições, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.6.1.3 – Continent

Por análise do gráfico abaixo, podemos verificar que a maioria dos países se situam na Europa (41), sendo que alguns se localizam em zonas próximas da Europa, nomeadamente na Ásia e em Africa. Existe ainda o caso da Austrália, que foi convidada a participar em 2015 e desde então tem participado todos os anos.

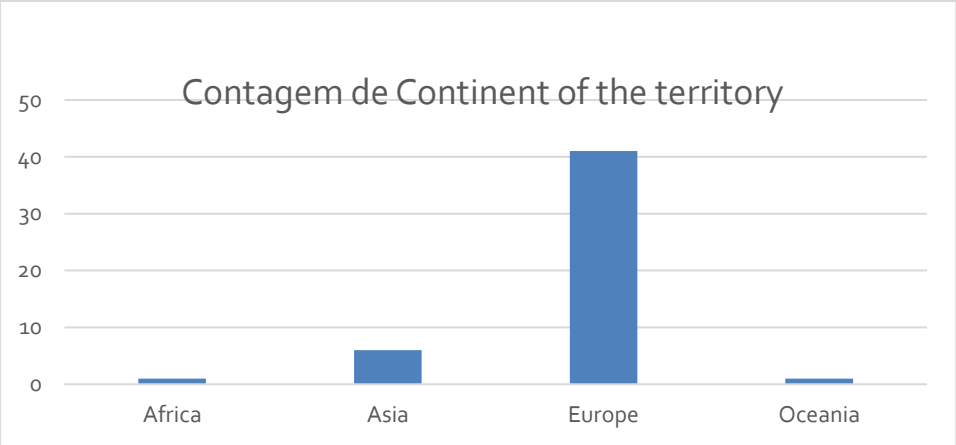

Figura 13 - Distribuição de valores do campo "Continent" do dataset Localização

2.6.1.4 – Region

Por análise do gráfico abaixo, podemos verificar que a região com mais participantes é a Europa do Sul (13).

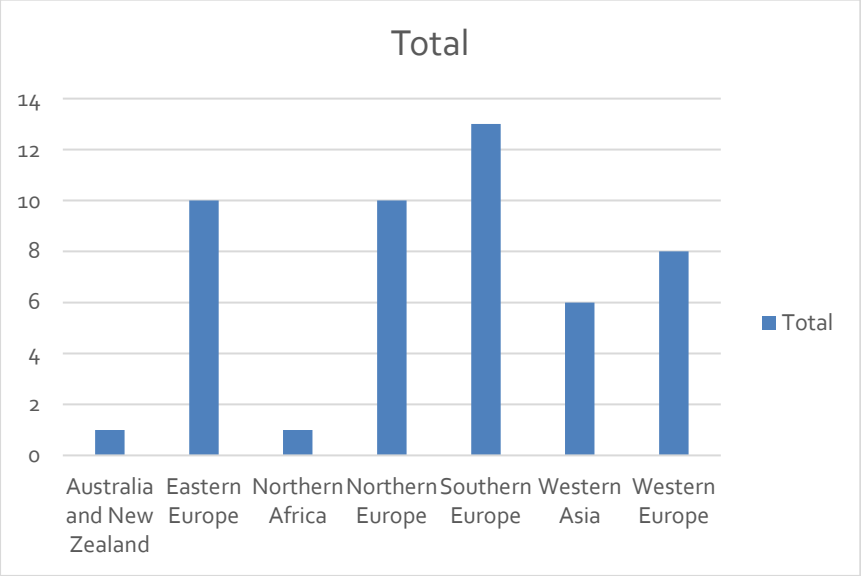

Figura 14 - Distribuição de valores do campo "Region" do dataset Localização

2.6.2 Erros e dados em falta

Não foram detetados erros nesta tabela.

2.7 Turistas

O ficheiro "*Turistas.xlsx*", após o download, continha os campos descritos na tabela seguinte.

Tabela 43 - Descrição dos campos do dataset Turistas

#	Campo	Tipo de dados	Descrição	Exemplo
1	Geo (labels)	Texto	País de destino dos turistas	Belgium
2	Time	Número	Ano a que se refere	2012
3	Number	Número	Número de turistas	757062

2.7.1 Análise estatística

A análise desta tabela consistiu principalmente em avaliar os valores da medida numérica, bem como a extensão dos dados.

2.7.1.1 – Geo (labels)

O campo "*Geo (labels)*" possui apenas a moda, que vai corresponder ao maior numero de campos não nulos, ou seja, os países que têm dados sobre o turismo no maior número de anos.

Tabela 44- Descrição do campo "Geo (labels)" do dataset Turistas

Campo	Moda	Ocorrências
Geo (labels)	Bulgária	30
	Alemanha	30
	Espanha	30
	Itália	30
	Luxemburgo	30
	Países Baixos	30
	Áustria	30
	Portugal	30
	Roménia	30
	Eslováquia	30

2.7.1.2 – Time

Analisando o campo "*Time*", conseguimos perceber o intervalo de tempo do dataset, visível na tabela seguinte.

Tabela 45 - Descrição do campo "Time" do dataset Turistas

Campo	Mínimo	Máximo
Time	1990	2019

2.7.1.3 – Number

Analisando o campo *"Number"*, é possível verificar que existe uma grande dispersão nos valores de turismo, como é possível ver na tabela e no histograma seguintes.

Tabela 46 - Descrição do campo "Number" do dataset Tourists

Campo	Mínimo	Máximo	Média	Desvio Padrão
Number	56 666	67 728 098	9 073 009	12798542.31

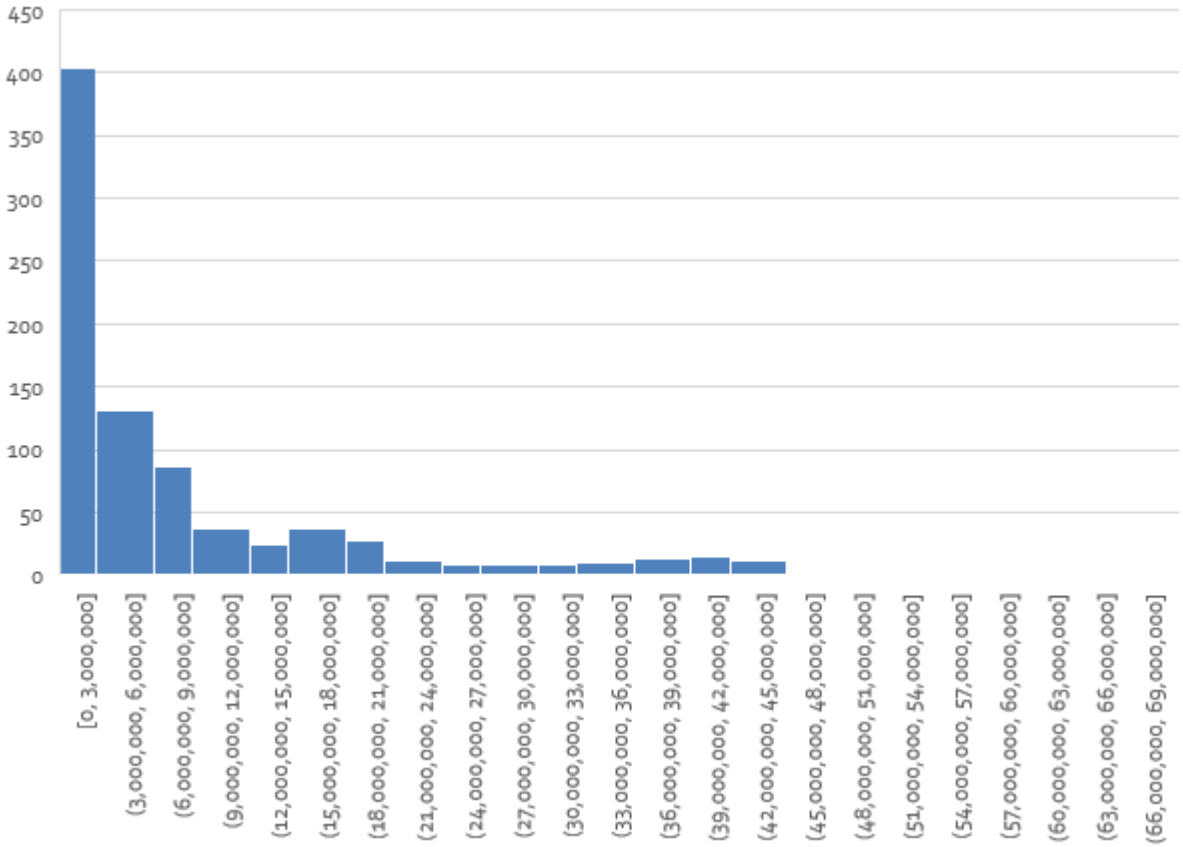

Figura 15 - Histograma número de turistas

2.7.2 Erros e dados em falta

A tabela tem em falta dados de alguns anos para alguns países e inclui apenas países da Europa, o que significa que existem países em falta. Além disso, alguns dos dados existentes na tabela contêm o indicativo de que se referem a estimativas ou valores com pouca precisão.

2.8 Emissões

O ficheiro *"Emissoes.xlsx"* guarda os dados de emissão de CO_2 de cada país no mundo. Este documento é constituído pelos seguintes campos.

Tabela 47 - Descrição dos campos do dataset Emissões

#	Campo	Texto	Descrição	Exemplo
1	Country	Texto	Nome do país	Italy
2	Gas	Texto	Tipo de gás emitido	CO2
2	Unit	Texto	Unidade de medida	$MtCO_2e$
3	Year	Número	Ano a que se refere a emissão	2000
4	Dados anuais	Número	Valores de gás emitido por ano	4749.57

Dado que este ficheiro apresenta dados de todos os países no mundo (195) e visto que o objeto de estudo são apenas os países que participaram/participam na eurovisão, eliminámos os dados referentes aos países que não integraram o concurso.

2.8.1 Análise estatística

2.8.1.1 – Country

O campo "*Country*" é um campo de texto que não apresenta moda visto que existem dados para todos os anos de praticamente todos os países na lista.

2.8.1.2 – Gas e Unit

Os campos "*Gas*" e "*Unit*" nesta tabela têm sempre o valor "CO_2" e "$MtCO_2$".

2.8.1.3 – Year

Analisando o campo "*Year*", conseguimos perceber o intervalo de tempo do dataset, visível na tabela seguinte.

Tabela 48 - Descrição do campo "Year" do dataset Emissoes

Campo	Mínimo	Máximo
Year	1990	2019

2.8.1.4 – Dados Anuais

À semelhança do dataset anterior, realizámos, novamente, a análise estatística da média, do valor máximo, do valor mínimo e do desvio padrão de emissões, sendo estas medidas ainda divididas por ano e por país.

Tabela 49 - Descrição do campo "Dados Anuais" do dataset Emissoes

Campo	Mínimo	Máximo	Média	Desvio Padrão
Dados anuais	0.14	1790.34	118.29	207.17

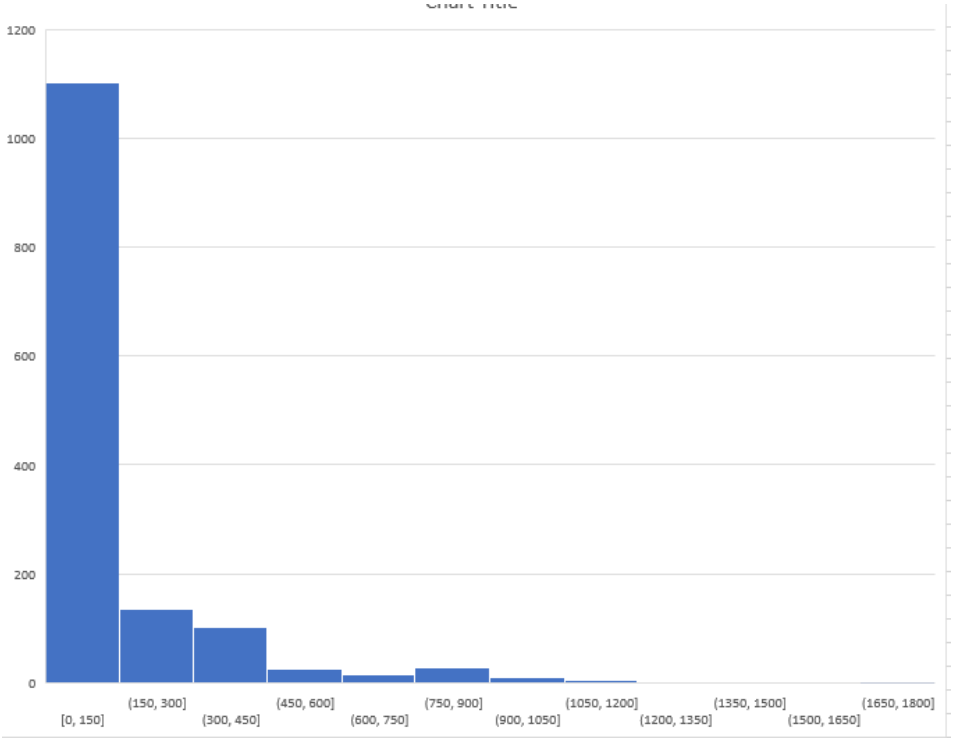

Figura 16 - Histograma do campo "Number" por país

Ao observar a figura 16, é visível que a maioria dos países emite valores de $C0_2$ baixos existindo alguns outliers.

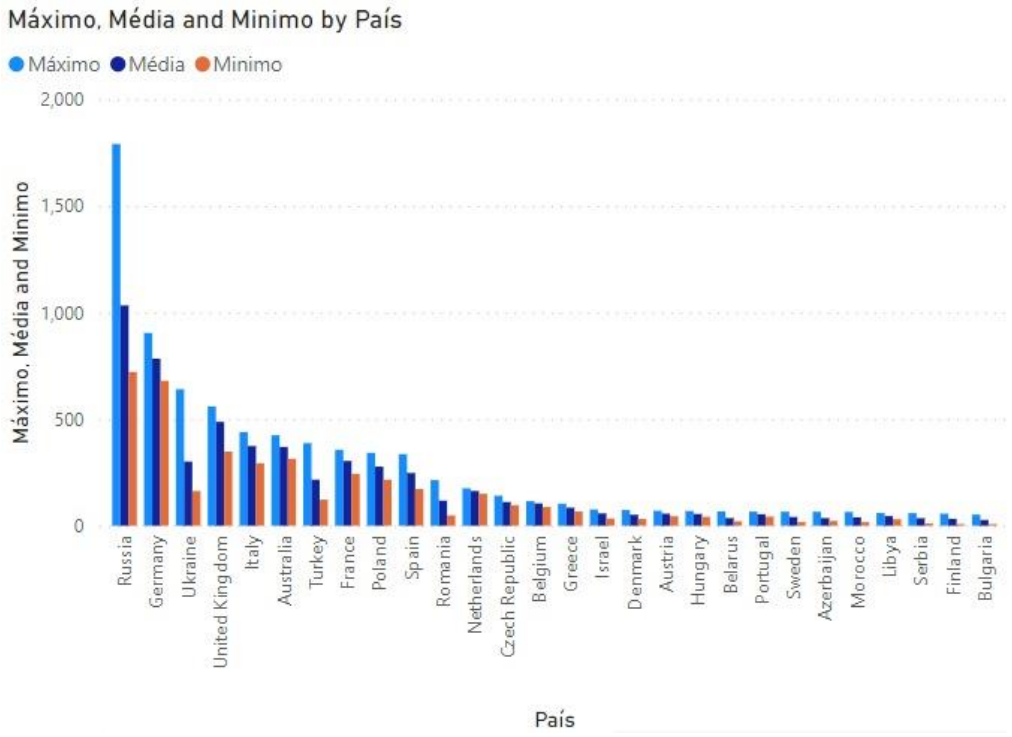

Figura 17 - Análise do campo "Number" por país

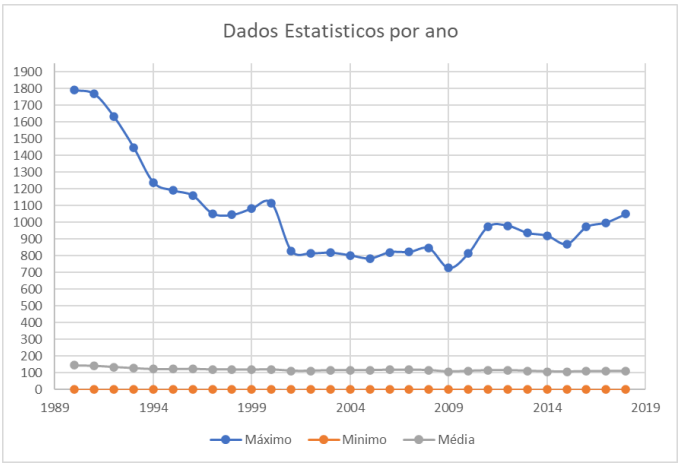

Figura 18 - Análise do campo "Number" por ano

Ao observar a figura "Dados estatísticos por ano" conseguimos determinar que o ano em que houve mais emissões de CO_2 foi 1990. De seguida começou a haver uma descida drástica das emissões até ao ano 2000, existindo um aumento das emissões nesse mesmo ano. Por fim, observa-se que as emissões de CO_2 tendem a apresentar um número constante, existindo ligeiras subidas e descidas ao longo do tempo.

Já na figura "Máximo, Média e Mínimo por País" é visível que o país da Eurovisão que emite mais CO_2 é a Rússia e que o país que emite menos é Andorra.

É possível realizar mais análises com este conjunto de dados, nomeadamente na variação das emissões de Co_2 em Portugal ao longo do tempo e um histograma das emissões de $C0_2$ de todos os países. Tal é possível observar nos dois gráficos seguintes.

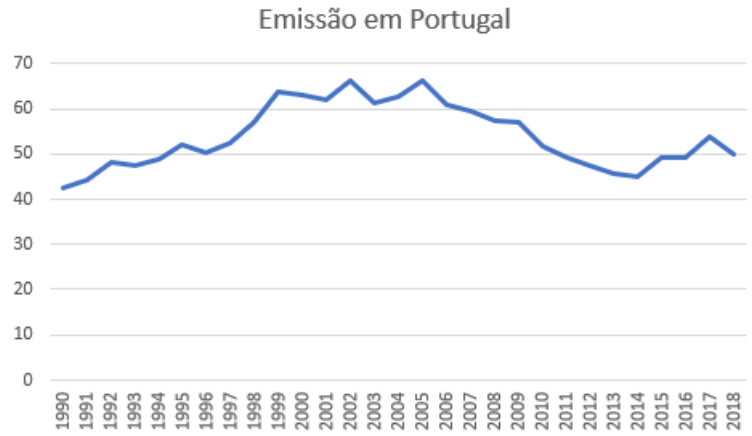

Figura 19: Emissões em Portugal por ano

2.8.2 Erros e dados em falta

Reparámos ao realizar a análise dos dados que existiam valores negativos. Em certos casos podem existir emissões de CO_2 negativas, no entanto nos casos considerados esses valores negativos constituíam *outliers* nos dados históricos e é mais provável que estes sejam devido a erro, pelo que corrigimos esses valores para que não houvesse contaminações nas observações.

2.9 PIB

O dataset extraído, referente ao Produto Interno Bruto (PIB), incluía 6 ficheiros no formato *.csv com uma estrutura igual à definida na tabela seguinte. Sendo a única diferença entre as tabelas a variável em estudo. Inicialmente tínhamos o valor do PIB, do crescimento do PIB, do PIB per capita, do crescimento do PIB per capita, do PIB-Paridade do Poder de Compra e do crescimento do PIB-Paridade do Poder de Compra (PPC), porém depois de uma breve ponderação, decidimos que a variável que iremos levar para as próximas fases do trabalho seria o valor do PIB per capital. Em cada um dos ficheiros temos dados anuais desde 1960 até 2020.

Tabela 50 - Descrição dos campos do dataset PIB

#	Campo	Tipo de dados	Descrição	Exemplo
1	Country	Texto	Nome do País ou da Região	Portugal
2	Year	Número	Ano a que se refere o valor do PIB per capita	2020
3	PIBpercapita	Número	Valor do PIB per capita	22,176.3 €

2.9.1 Análise estatística

2.9.1.1 Country

O campo "*Country*" é um campo de texto único, sem repetições, pelo que não faz sentido realizar qualquer análise estatística sobre este.

2.9.1.2 Year

Tabela 51-Descrição do campo "Year" do dataset PIB

Campo	Mínimo	Máximo
Year	1960	2020

Os dados variam temporalmente entre 1960 e 2020.

2.9.1.3 PIB per Capita

Tabela 52 - Descrição do campo "PIBpercapita" do dataset PIB

Campo	Mínimo	Máximo	Média	Desvio Padrão
PIBpercapita	60.45821	189487.1	19041.34	24369.11

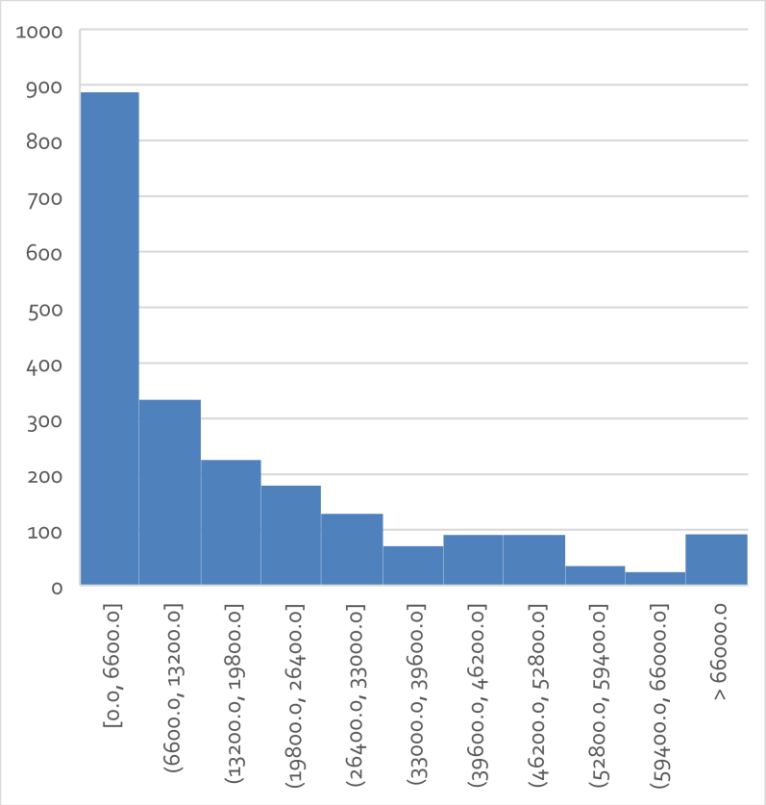

Figura 20-Histograma do PIB per capita de 1960 a 2020

2.9.2 Erros e dados em falta

Como seria de esperar, muitos dos países que participaram da Eurovisão não apresentam dados desde 1960 pelo que, posteriormente e em futuras análises, poderão existir lacunas de informação em algumas bandas temporais.

2.10 População

A tabela "*Populacao.xlsx*" contêm o número de população total anual para uma lista de países e regiões do mundo de 1960 a 2020. No âmbito deste trabalho iremos considerar apenas a população residente em países da Eurovisão.

Tabela 53 - Descrição dos campos do dataset Populacao

#	Campo	Tipo de dados	Descrição	Exemplo
1	Country	Texto	Nome do País ou da Região	Portugal
2	Year	Número	Ano a que se refere o valor da população total	2020
3	Population	Número	Valor da população total	10,305,564

2.10.1 Análise estatística

2.10.1.1 – Country

O campo "*Country*" é um campo de texto que não apresenta moda visto que existem dados para todos os anos de praticamente todos os países na lista.

2.10.1.2 - Year

Tabela 54 - Descrição do campo "Year" do dataset Populacao

Campo	Mínimo	Máximo
Year	1960	2019

Os dados variam temporalmente entre 1960 e 2019.

2.10.1.3 - Population

Tabela 55 - Descrição do campo "População" do dataset Populacao

Campo	Mínimo	Máximo	Média	Desvio Padrão
População	13410	148538197	16327054.5	26050378.1

Existe uma grande dispersão de valores de população pelos países da Eurovisão, o que é expectável visto que participaram no festival desde microestados com baixa população como Andorra, San Marino ou Mónaco até às maiores nações europeias, como Alemanha, França e Reino Unido.

Nos gráficos seguintes observa-se a variação da população ao longo do tempo nos países da Eurovisão e também um histograma com a sua distribuição.

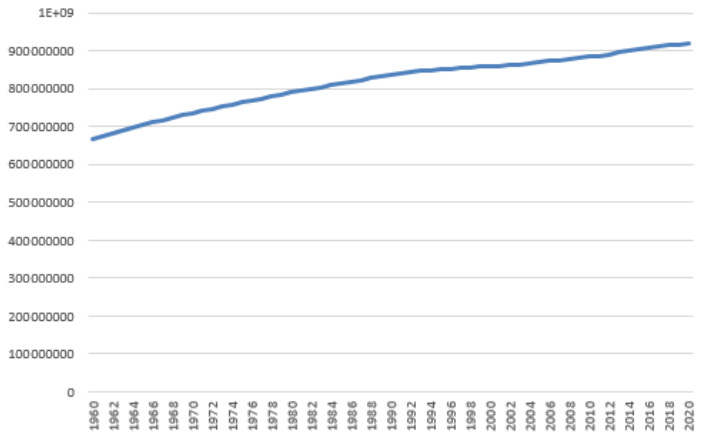

Figura 21 - Crescimento populacional na área Eurovisão

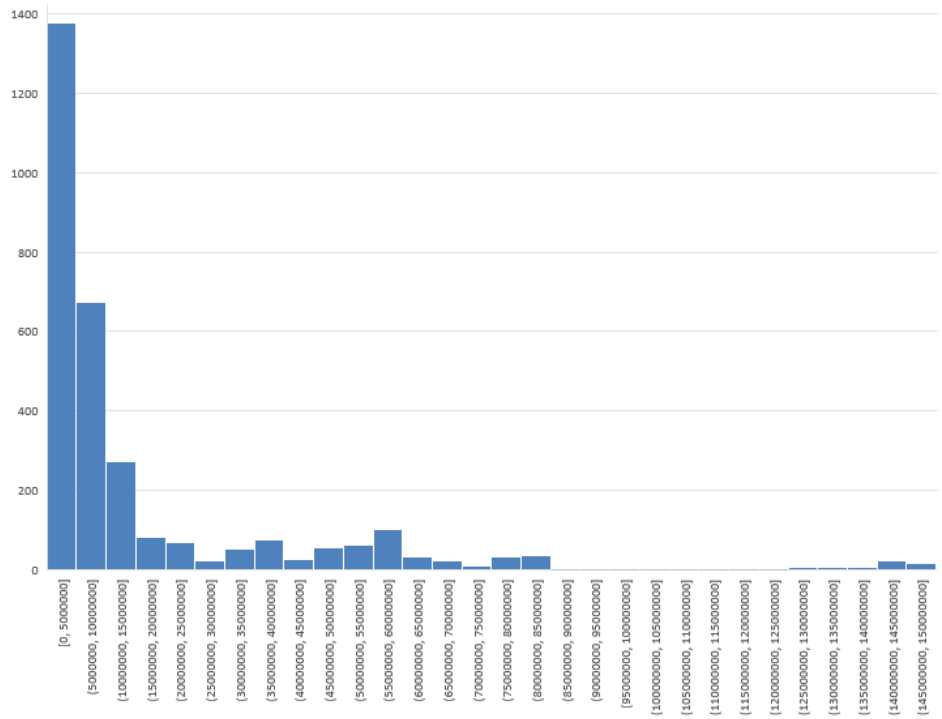

Figura 22 - Histograma de população nos países da área Eurovisão

2.10.2 Erros e dados em falta

Não foram detetados erros nem lacunas nos dados.

2.11 Conflitos

O ficheiro *"conflitos.xlsx"* é uma lista de todos os conflitos em que pelo menos uma nação soberana esteve envolvida. Para cada conflito temos - o nome da guerra, a data de início e término e as nações que lutaram nela, tal como descrito abaixo.

Tabela 56 - Descrição dos campos do dataset Conflitos

Campo	Texto	Descrição	Exemplo
Conflict	Texto	Nome do conflito	Gulf War
Start Date	Data	Data de início	1990-08-02T00:00:00Z
End Date	Data	Data de fim	1996-10-24T00:00:00Z
Participant1 **(…)** **Participant20**	Texto	Participantes no conflito	United Kingdom

Devido a erros detetados na análise, descritos na secção 2.11.2, foi criada uma nova tabela sobre a qual irá recair a análise estatística.

Tabela 57 - Descrição dos novos campos da tabela do dataset Conflitos

#	Campo	Texto	Descrição	Exemplo
1	ID	Categórico	Identificador único	1
2	Conflict Location	Texto	País ou local onde o conflito se realizou maioritariamente	Morocco

3	EurovisionCountry	Categórico	Campo que verifica se a localização do conflito pertence ou não a um país participante na Eurovisão	"EurovisionParticipant"
4	Conflict Name	Texto	Nome do Conflito	Ifni War
5	Start Date	Data	Dada de início	23/10/1957
5	End Date	Data	Data de fim	30/06/1958
6	Participant	Texto	Nome do país participante	Spain

É de notar que nesta tabela Excel constam apenas os dados desde o início do festival da Eurovisão e apenas aqueles em que os países participantes na Eurovisão foram participantes ativos.

2.11.1 Análise Estatística

2.11.1.1 – ID

Tabela 58 - Descrição do campo "ID" do dataset Conflitos

Campo	Max	Min
ID	185	1

Existem no dataset 185 instâncias de países da Eurovisão a participar em conflitos, o que não corresponde a 185 conflitos, mas sim uma linha por país por conflito.

2.11.1.2 - Conflict Location

O valor mais comum para localização do conflito é a Faixa de Gaza onde existem 7 ocorrências

Tabela 59 - Descrição do campo "Conflict Location" do dataset Conflitos

Campo	Moda	Ocorrências
Conflict Location	Gaza Strip	7

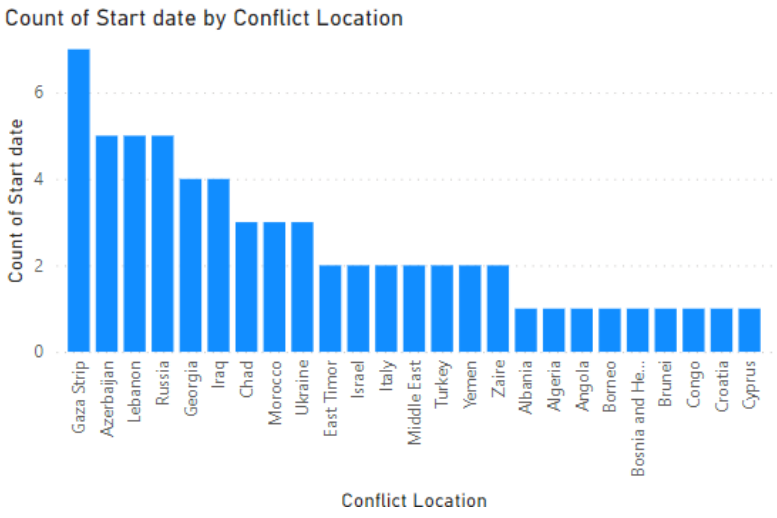

Figura 23 - Localizações com maior ocorrência de conflitos no dataset Conflicts Participants

2.11.1.3 – EurovisionCountry

O campo "*EurovisionCountry*" é um campo de verificação que assume valores "*EurovisionParticipant*" ou "*DoesNotParticipateInEurovision*".

Tabela 60 - Descrição do campo "EurovisionCountry" do dataset Conflitos

Campo	Valor	Ocorrências
EurovisionCountry	EurovisionParticipant	75
	DoesNotParticipateInEurovision	110

Na tabela existem 75 ocorrências de países a participar em conflitos no solo de países da Eurovisão e 110 instâncias de participação em conflitos fora da área da Eurovisão.

2.11.1.4 – Conflict Name

O campo "*Conflict Name*" é um campo de texto único. Embora existam conflitos que tiveram várias partes, estes normalmente aparecem na tabela com diferentes nomes, numerados ou com data associada no título, por exemplo "*Shaba I*" e "*Shaba II*" ou "*2016 Armenian–Azerbaijani clashes*" e "*2018 Armenian–Azerbaijani clashes*" pelo que não faz sentido realizar qualquer análise estatística sobre este campo.

2.11.1.5 - Start Date e End Date

As datas de início e final do conflito variam entre 1957 e 2022, sendo que existem conflitos na tabela que não foram ainda concluídos e que vão ter o valor de "*Ongoing*"

Tabela 61 - Descrição dos campos "Start Date" e "End Date" do dataset Conflitos

Campo	Max	Min
Start date	24/02/2022	23/10/1957
End date	21/05/2021	30/06/1958

2.11.1.6 – Participant

Relativamente aos participantes, o país que participou em mais conflitos durante este período de tempo foi a França, com um total de 18.

Tabela 62 - Descrição do campo "Participant" do dataset Conflitos

Campo	Moda	# Ocorrencias
Participant	France	18

De notar no caso dos participantes, a Rússia, no segundo lugar com participação em 17 conflitos, conta ainda com 3 conflitos na lista durante os anos da União Soviética.

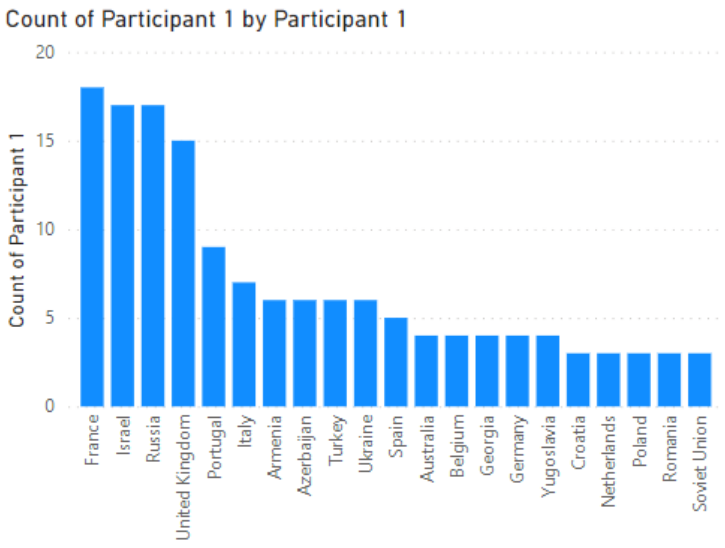

Figura 24 - Países com mais participações em conflitos no dataset Conflicts Participants

2.11.2 Erros e dados em falta

Ao analisar as tabelas foi descoberto que algumas datas de início e fim são iguais para os conflitos. Os dados foram corrigidos para os conflitos que irão ser utilizados no trabalho, nomeadamente os que envolvam países que participam na Eurovisão e que aconteceram desde 1957. De forma a colmatar as lacunas nos dados e a corrigir os erros detetados inicialmente, este dataset foi adaptado, no entanto durante a pesquisa para colmatar as falhas nas datas foram ainda encontradas mais instâncias de conflitos, com recurso a fontes de informação online (Wikipedia, 2022), tendo por isso sido criada uma tabela Excel que aglomera os dados originais com os dados obtidos durante esta pesquisa.

2.12 Área

Durante a segunda fase do projeto sentimos necessidade de normalizar alguns dados de outros datasets devido à grande dispersão de valores, por exemplo, de emissões de dióxido de carbono, que não têm em conta outros fatores e por isso comparam linearmente países tão diferentes como a Islândia e a Rússia. A forma que encontrámos de tentar normalizar esta informação foi através da utilização dos valores de população do país, mas também da área que um país ocupa. Assim sendo, acrescentámos mais um dataset ao projeto.

A tabela "*Area.xlsx*", tal como importada, incluía dois campos que continham um INDICATOR_CODE e INDICATOR_NAME com informação redundante por isso procedemos à eliminação das mesmas. Esta tabela, retirada do World Bank, expõe a área terrestre de todos os países e regiões do mundo.

É importante referir que por área terrestre entende-se a área total de um país, excluindo a área sob corpos de água interiores, reivindicações nacionais à plataforma continental e zonas económicas exclusivas. Este dataset inclui ainda as variações de área ao longo do tempo devido à formação de novos países e áreas disputadas.

Dado que a organização da tabela original não se adaptava aos objetivos do trabalho, alterámos a disposição dos campos e reduzimos o volume de dados de forma a manter só os dados relevantes para o projeto, nomeadamente de países participantes na Eurovisão. O resultado encontra-se descrito na tabela abaixo.

Tabela 63 - Descrição dos campos do dataset Área

#	Campo	Tipo de dados	Descrição	Exemplo
1	Country	Texto	Nome do País ou da Região	Portugal
2	Year	Número	Ano a que se refere o valor numérico com a área	1961
3	LandArea	Número	Área total do país, medido em quilômetros quadrados (km²)	91500

2.12.1 Análise Estatística

2.12.1.1 – Country

O campo "*Country*" é um campo de texto que não apresenta moda visto que existem dados para todos os anos (61) de praticamente todos os países na lista, à exceção do Luxemburgo.

2.12.1.2 – Year

Tabela 64 - Descrição do campo "Year" do dataset Area

Campo	Mínimo	Máximo
Year	1961	2021

Os dados da tabela variam entre 1961 e 2021 para os países selecionados.

2.12.1.3 – LandArea

Tabela 65 - Descrição do campo "LandArea" do dataset Area

Campo	Mínimo	Máximo	Média	Desvio Padrão
LandArea	2.027	16389950	648956.77	2507791.29

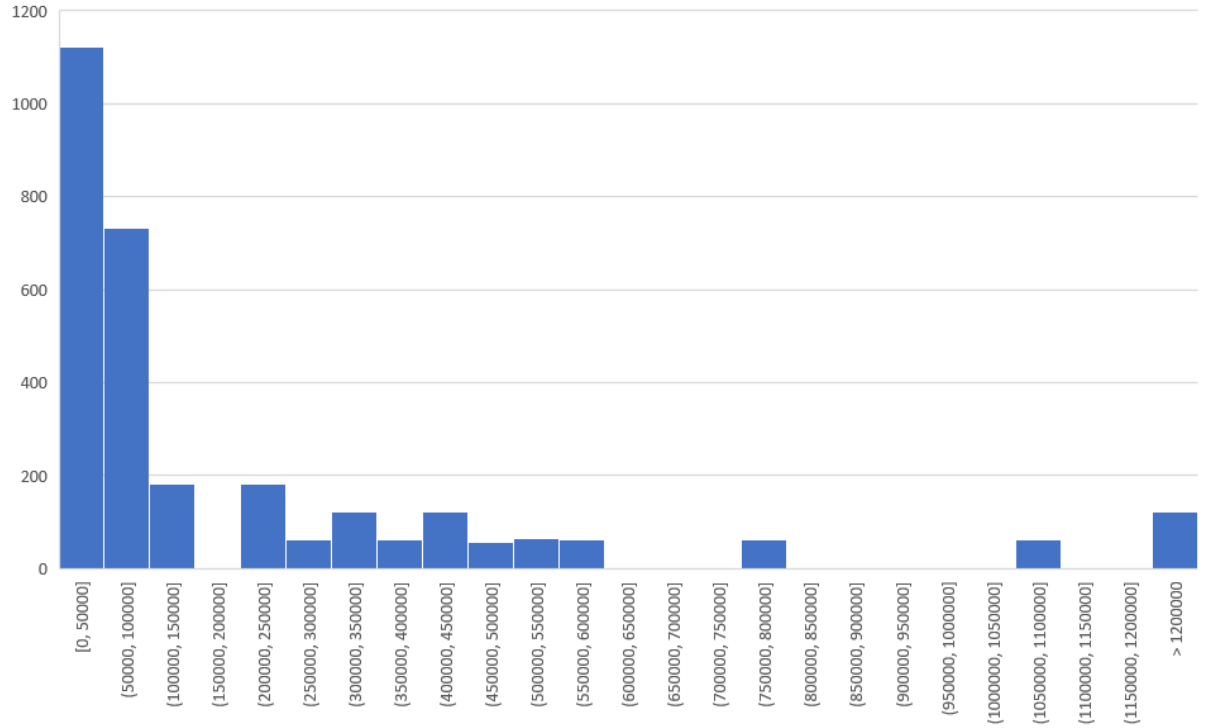

Figura 25 - Histograma do campo "LandArea" do dataset Area

Como é visível pela tabela e gráfico anteriores, existem muitos países com baixa área, alguns países com áreas médias e poucos países com áreas muito elevadas, o que aumenta a média e o desvio padrão dos dados das zonas consideradas.

2.12.2 Erros nos dados

Não foram detetados erros nos dados, no entanto existem alguns valores em falta, nomeadamente para o Luxemburgo entre 1961 e 1999.

3. Diagrama Relacional das Fontes de Dados

Com o intuito de observar as ligações entre cada conjunto de dados, elaborou-se o seguinte esquema.

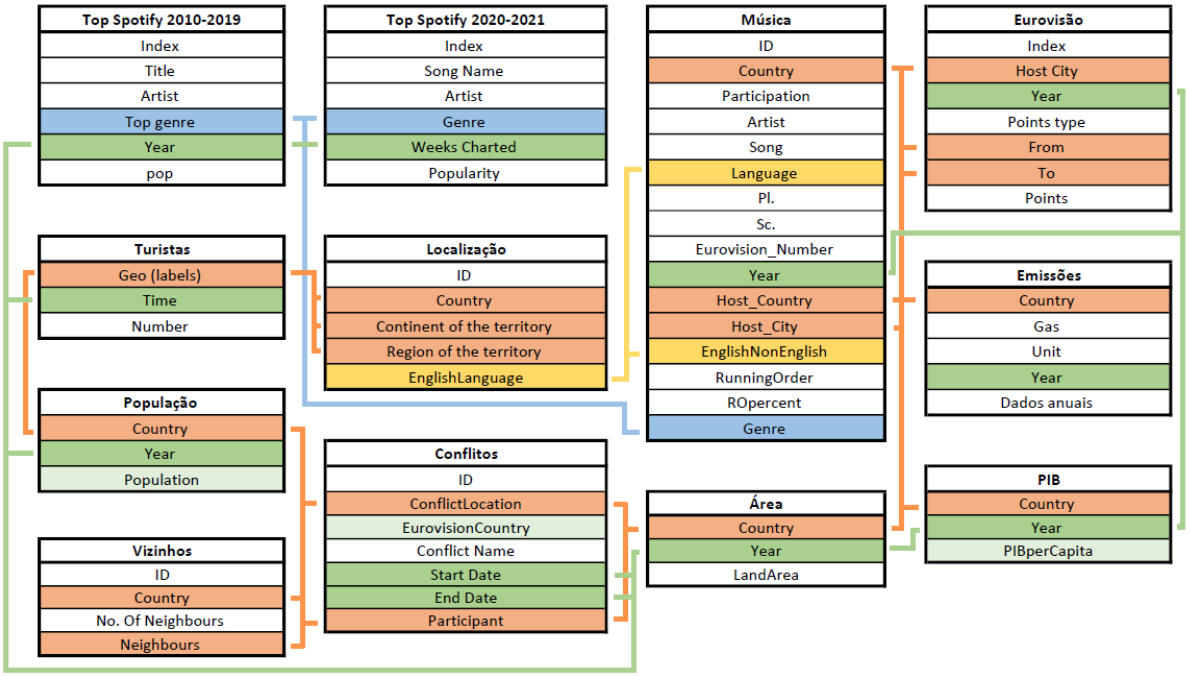

Figura 26 - Diagrama relacional entre tabelas

Neste, realizaram-se várias ligações entre os diversos datasets, unindo os campos com o mesmo tipo de dados.

Foram criadas quatro cores para relacionar o conteúdo dos campos:

- A cor verde refere-se ao espaço temporal – Data, Ano
- A cor azul refere-se ao género musical
- A cor laranja refere-se aos países – Localização
- A cor amarela refere-se à língua da música.

4. Processo de Negócio

Define-se processo de negócio como um conjunto de atividades ou tarefas estruturadas relacionadas que produzem um serviço ou produto específico para os seus clientes ou para um cliente particular.

Aplicando a definição teórica anteriormente descrita ao nosso caso de estudo, podemos definir duas possíveis entidades que poderiam ter interesse em utilizar os dados recolhidos que e que irão ser analisados:

- Países
- Casas de apostas

Relativamente aos países, fará parte do objetivo deste projeto, também refletido nas questões analíticas, a relação entre um país ganhar a eurovisão e o possível aumento do PIB e do turismo no ano seguinte. Tal poderá ser interessante a um país que tenha interesse em ganhar a Eurovisão para usufruir destes acréscimos e utilizar as outras análises dos dados no que diz respeito, por exemplo, aos géneros e às linguagens de música que obtém melhores resultados, concorrendo, assim, ao festival com músicas que apresentem uma melhor receção pela parte do público.

Por outro lado, as casas de apostas permitem aos seus utilizadores fazer apostas monetárias sobre vários tipos de categorias, como o desporto, as eleições e outros eventos da atualidade. Uma casa de apostas consegue obter lucro através da cobrança de comissões por cada aposta realizada. Esta comissão é cobrada através de uma ligeira manipulação das probabilidades e pagamentos de cada aposta.

Um exemplo prático seria uma aposta de lançamento de moeda – existem dois resultados possíveis (cara ou coroa), então a probabilidade de cada uma destas acontecer será 50%, o que resulta numa probabilidade decimal de 2.00 (ou, na prática, o lucro para a pessoa que fez a aposta será o dobro da aposta original). Na realidade, o que acontece é que a casa de apostas estabelece estas probabilidades decimais ligeiramente abaixo das reais, por exemplo 1.9 para o caso acima, sendo a diferença entre elas o lucro para a casa (Reyes, 2022).

As casas de apostas necessitam de ter cuidado quando estabelecem as probabilidades base para qualquer tipo de jogo, de forma a balançar a atratividade para os utilizadores e o risco para a companhia. Os casos de apostas reais não são normalmente tão simples como o exemplo acima referido, existindo durante o tempo em que uma aposta esta aberta um ajuste das probabilidades.

O nosso projeto poderá ser utilizado para gerar essas percentagens base com fundamento em dados históricos e tendências de voto do público, assim como para fazer atualizações às probabilidades durante o decorrer das semanas próximas do festival, em que dados como a ordem das músicas no concurso são publicados.

		winning chance	BET365	SMARKETS*	BETSSON
1	Ukraine Kalush Orchestra - Stefa...	28%	2.37	2.94	2.3
2	Italy Mahmood & Blanco - Brividi	18%	3.5	3.8	3.95
3	Sweden	6%	12	15	11
4	Greece artist: Amanda Tenfjord	5%	12	15	13
5	Poland Krystian Ochman - River	4%	15	24	18
6	United Kingdom	4%	17	20	17
7	Norway Subwoolfer - Give That Wolf a Ban...	3%	23	48	20
8	Netherlands S10 - De Diepte	3%	26	32	25
9	Belgium artist: Jérémie Makiese	2%	26	48	25
10	Spain Chanel - SloMo	2%	34	44	30
11	Australia Sheldon Riley - Not the Same	2%	29	48	30
12	France Alvan & Ahez - Fulenn	2%	36	44	35
13	Switzerland Marius Bear - Boys Do Cry	2%	21	160	30
14	Finland The Rasmus - Jezebel	2%	51	100	40
15	Cyprus Andromache - Ela	2%	41	65	50

		winning chance	BET365	UNIBET	888 SPORT
1	Ukraine Kalush Orchestra - Stefa...	33%	6/5	6/5	6/7
2	Italy Mahmood & Blanco - Brividi	17%	10/3	13/4	3/1
3	Sweden Cornelia Jakobs - Hold Me Closer	12%	5/1	19/4	10/3
4	United Kingdom Sam Ryder - Space Man	5%	14/1	17/1	13/1
5	Greece Amanda Tenfjord - Die Together	4%	18/1	20/1	14/1
6	Poland Ochman - River	4%	18/1	20/1	14/1
7	Spain Chanel - SloMo	2%	25/1	33/1	28/1
8	Norway Subwoolfer - Give That Wolf a Ba...	2%	33/1	30/1	32/1
9	Netherlands S10 - De Diepte	2%	33/1	30/1	27/1
10	Australia Sheldon Riley - Not the Same	2%	40/1	50/1	39/1
11	Portugal Maro - Saudade, saudade	1%	50/1	50/1	45/1
12	Belgium Jérémie Makiese - Miss You	1%	66/1	50/1	50/1
13	France Alvan & Ahez - Fulenn	1%	66/1	50/1	66/1
14	Switzerland Marius Bear - Boys Do Cry	1%	66/1	66/1	66/1
15	Serbia Konstrakta - In Corpore Sano	1%	66/1	60/1	74/1

		winning chance	888 SPORT	BET365	UNIBET	LAD BROKES	SMARKETS*	COOL BET
1	Ukraine Kalush Orchestra - Stefa...	42%	1.6	1.8	1.6	1.73	1.92	1.93
2	Italy Mahmood & Blanco - Brividi	15%	5.5	5	5	5	6.6	6
3	Sweden Cornelia Jakobs - Hold Me Closer	11%	6.7	7	6.5	7	8.8	6.5
4	United Kingdom Sam Ryder - Space Man	7%	9	11	13	11	14	11
5	Spain Chanel - SloMo	4%	17	17	19	19	24	22
6	Greece Amanda Tenfjord - Die Together	3%	22	15	31	26	34	28
7	Poland Ochman - River	2%	41	41	26	34	75	51
8	Norway Subwoolfer - Give That Wolf a B...	2%	43	41	31	41	80	51
9	Netherlands S10 - De Diepte	1%	50	51	51	41	120	71
10	France Alvan & Ahez - Fulenn	1%	77	67	67	41	100	71
11	Australia Sheldon Riley - Not the Same	1%	70	71	81	67	130	81
12	Portugal Maro - Saudade, saudade	1%	73	71	81	67	120	71
13	Serbia Konstrakta - In Corpore Sano	1%	95	101	81	81	140	91
14	Finland The Rasmus - Jezebel	1%	140	126	81	101	160	101
15	Switzerland Marius Bear - Boys Do Cry	1%	106	126	81	126	230	201

Figura 27 - Probabilidade de vencer a Eurovisão em 9/03/2022, 31/03/2022 e 01/05/2022 (EurovisionWorld, 2022)

5. Questões Analíticas

Com o objetivo de concretizar os elementos que serão estudados na fase seguinte do projeto, foram elaboradas as seguintes questões analíticas:

1. Qual a influência da língua em que a canção é cantada? Existe maior quantidade de países que não se qualificam para a final cuja língua da música não seja o inglês? Existe melhor resultado médio para músicas em inglês? Existe alguma diferença entre os resultados do mesmo país entre músicas em inglês ou com a sua língua materna?
2. Como é que a demografia e a geografia influenciam os resultados na eurovisão? Um país tem influência na quantidade de pontos que recebe? Existe entreajuda entre vizinhos? Os países com maior população vizinha têm vantagens? Os países com maior PIB têm melhores resultados?
3. As questões da atualidade influenciam os resultados? Existe correlação entre os géneros musicais mais ouvidos em cada ano e a Eurovisão? A participação em conflitos diminui a média de pontos que um país recebe? Os países mais "verdes" são mais populares? O turismo influencia a votação?

6. Modelação Dimensional

6.1 Declaração do grão e tipo da tabela de factos

A tabela de factos é a tabela principal dentro de um modelo multidimensional do tipo Star Schema, criado por Ralph Kimball. Esta tem como característica principal uma elevada quantidade de dados redundantes para se obter um melhor desempenho.
Dentro da tabela de factos cada facto é normalmente identificado por uma chave composta – constituída por várias chaves estrangeiras - que pode ser associado ao grão. Entende-se grão como o significado de cada linha da tabela de factos, estando este relacionado com o nível máximo de detalhe. Quanto mais fino for o grão, maior o número de dimensões, ou seja, maior o número de atributos da chave estrangeira da tabela de factos.
No presente trabalho tivemos a necessidade de criar duas tabelas de factos com granularidade diferentes.
Na primeira tabela de factos cada linha da tabela corresponderá ao número de pontos (valor numérico), que um país A dá a um país B, num determinado Ano (em que se realizou a Eurovisão), havendo também a indicação do tipo de pontos que é dado (podendo estes serem dados pelo publico ou um júri).
Na segunda tabela de factos, cada linha corresponderá ao número de pontos (valor numérico) que uma música apresentada por um país A, num determinado Ano (em que se realizou a Eurovisão) recebe.
Uma vez que os grãos de ambas as tabelas consistem numa linha por transação, podemos considerar que temos perante uma tabela de factos do tipo transacional.
A tabela de factos, que se encontra no centro do esquema, está rodeada pelas tabelas de dimensão. A primeira tabela armazena grande quantidade de dados históricos, em função do tempo, que correspondem a cada instância em que um país dá pontos a outro país. Para esta tabela de factos foi considerado um menor número de dimensões, pois o âmbito é apenas de descrever a relação entre os dois países que integram cada linha. A segunda tabela apresenta dados que representam uma visão agregada da performance de cada música, ou seja, de um país num determinado ano. Os valores dos pontos para cada música podem ser obtidos por soma dos pontos que constam na primeira tabela de factos, no entanto, o âmbito desta tabela de factos será o de encontrar correlações positivas e negativas com outros acontecimentos anuais no país, sendo por isso necessário adicionar a esta tabela um conjunto de dados que não faz sentido ao nível da granularidade da primeira tabela.

6.1.1 – Dimensões na tabela de factos 1

Grão: No Ano A, o País B deu ao País C o número de Pontos X do Tipo D

Tabela 66 - Descrição das dimensões da tabela de factos 1

Campo	Descrição	Dimensão Origem
IDData	Chave Estrangeira	Data
IDEurovisao	Chave Estrangeira	Eurovisão
IDLocalizacaoDa	Chave Estrangeira	Localização
IDLocalizacaoRecebe	Chave Estrangeira	Localização
IDJunkDimension	Chave Estrangeira	Junk Dimension

6.1.2 – Dimensões na tabela de factos 2

Grão: Música (País A recebeu o número de Pontos X no Ano B)

Tabela 67 - Descrição das dimensões da tabela de factos 2

Campo	Descrição	Dimensão Origem
IDData	Chave Estrangeira	Data
IDEurovisao	Chave Estrangeira	Eurovisão
IDLocalizao	Chave Estrangeira	Localização
IDMusica	Chave Estrangeira	Música
IDGrupoConflito	Chave Estrangeira	GrupoConflito

6.2 Tabelas de Dimensão

6.2.1 Dimensão Localização

Hierarquia identificada apresenta profundidade fixa:
Continente > Região > País

Tabela 68 - Descrição da Dimensão Localização

Campo	Descrição	Origem dos dados	Tipo de dados	Valores	[1]
IDLocalizacao	Chave substituta	ID único gerado sequencialmente	C	1-49	-
País	País a que se refere o ID	Tabela Localização campo Country	T	Ex: Portugal	-
Região	Região a que se refere o País	Tabela Localização campo Region	T	Ex: Eastern Europe	-
Continente	Continente a que se refere o País	Tabela Localização campo Continent	C	Europa, Asia, Oceânia	-
Língua	Língua do país	Tabela Localização campo Língua	C	English, Not_English, Mixed	-
Número Vizinhos	Número de vizinhos que cada país	Tabela Localização campo No. Of Neighbours	N	0-16	-

[1] Valor Inserido quando desconhecido ou não aplicável

Visualização exemplificativa

IDLocalizacao	País	Continente	Região	Língua	Número Vizinhos
1	Albania	Europe	Southern Europe	Not_English	4
2	Andorra	Europe	Southern Europe	Not_English	2
3	Armenia	Asia	Western Asia	Not_English	4
4	Australia	Oceania	Australia and New Zealand	English	0
5	Austria	Europe	Western Europe	Not_English	8
6	Azerbaijan	Asia	Western Asia	Not_English	5
…	…	…	…	…	…

6.2.2 Dimensão Música

Esta dimensão não apresenta hierarquias.

Tabela 69 - Descrição da Dimensão Música

Campo	Descrição	Origem dos dados	Tipo de dados	Valores	[1]
IDMusica	Chave substituta	ID único gerado sequencialmente	C	1-646	-
NomeMusica	Música cantada na eurovisão	Tabela Eurovision song lyrics, campo Value.Song	T	Ex: Sangen om dig	-
Língua	Língua da música	Tabela Eurovision song lyrics, campo EnglishNonEnglish	C	English, Not_English, Mixed	-
Classificação	Classificação final da música	Tabela Eurovision song lyrics, campo ValuePl	N	1-26	NA (not applicable)
OrdemAtuacao	Número da ordem do concurso em que a música tocou	Tabela Eurovision song lyrics, campo RunningOrder	N	1-27	NQ (not qualified)
Percentagem OrdemAtuacao	Razão entre o RunningOrder e o número total de músicas tocadas na final	Tabela Eurovision song lyrics, campo Ropercent	N	0.0-1.00	NA (not applicable)
Pontuação	Pontuação final	Tabela Eurovision song lyrics, campo Value.Sc.	N	0-758	NA (not applicable)

[1] Valor Inserido quando desconhecido ou não aplicável

Na tabela acima, foram utilizados valores NQ quando as músicas não foram à final do concurso. Já o valor NA utilizou-se quando não houve pontuação e não houve atuação.

Visualização exemplificativa

IDMusica	Música	Lingua	Classificação	RunningOrder	Percentagem OrdemAtuacao	Pontuação
…	…	…	…	…	…	…
1556	Arcade	English	1	12	0.461538	498
1557	Proud	English	7	8	0.307692	305
1558	Truth	English	8	20	0.769231	302
1559	Sister	English	25	4	0.153846	24
1560	Home	English	23	14	0.538462	35
…	…	…	…	…	…	…

6.2.3 Dimensão Data

Hierarquia identificada apresenta profundidade fixa.
Século > Década > Parte da Década > Ano

Tabela 70 - Descrição da Dimensão Data

Campo	Descrição	Origem dos dados	Tipo de dados	Intervalo	[1]
IDData	Chave substituta	ID único gerado sequencialmente	C	1-64	-
Ano	Valor numérico correspondente ao calendário Gregoriano	Tabela Eurovision Final voting results, campo Year	N	1957-2021 Exceto 2020	-
Parte da Década	Valor texto que indica em que metade da década se encontra o Ano	Obtido através de funções	C	"Primeira Metade da Década" "Segunda Metade da Década"	-
Década	Valor numérico correspondente à década	Obtido através de funções	C	1950-2020	-
Século	Valor numérico correspondente ao século em que o ano gregoriano se encontra	Obtido através de funções	C	20-21	-

[1] Valor Inserido quando desconhecido ou não aplicável

Visualização exemplificativa

ID	Ano	Parte Da Década	Década	Século
1	1957	Segunda Metade da Década	1950	20
2	1958	Segunda Metade da Década	1950	20
3	1959	Segunda Metade da Década	1950	20
4	1960	Primeira Metade da Década	1960	20
5	1961	Primeira Metade da Década	1960	20
…	…	…	…	…

6.2.4 Dimensão Conflitos

Esta dimensão não apresenta hierarquias. A informação irá ser condensada numa tabela Dimensão GrupoConflito, que será descrita mais à frente.

Tabela 71 - Descrição da Dimensão Conflitos

Campo	Descrição	Origem dos dados	Tipo de dados	Valores	[1]
ID**Conflito**	Chave substituta	ID gerado sequencialmente	C	1 - 185	-
Conflict Location	País onde o conflito se realizou maioritariamente	Tabela Conflitos Campo Conflict Location	T	Ex: "Morocco"	-
Eurovision Country	Verifica se o conflito se localizou num território pertencente a um país da Eurovisão	Tabela Conflitos Campo EurovisionCountry	C	Ex. "EurovisionParticipant"	
Conflict Name	Nome do conflito	Tabela Conflitos Campo Conflict Name	T	Ex: "Ifni War"	-
Participant	Nome do país participante	Tabela Conflitos Campo Participant	T	Ex: "Spain"	-
Start Year	Ano de início	Obtido através de funções a partir do campo Start Date (tabela Conflitos)	N	1957-2022	-
End Year	Ano de Término	Obtido através de funções a partir do campo End Date (tabela Conflitos)	N	1958 – 2022	Ongoing
State	Indica a situação do conflito durante Eurovisão	Indicador gerado através de funções, tendo em conta as datas da Eurovisão	C	"Ativo" "Não Ativo"	-

[1] Valor Inserido quando desconhecido ou não aplicável

Visualização exemplificativa

IDConflito	**Eurovision Country**	**…**	**Participant**	**Start Year**	**End Year**	**State**
1	EurovisionParticipant	…	Spain	1957	1958	AtivoEurovisao
2	EurovisionParticipant	…	France	1957	1958	AtivoEurovisao
3	DoesNotParticipateIn Eurovision	…	Portugal	1959	1959	NãoAtivoEurovisao
4	DoesNotParticipateIn Eurovision	…	United Kingdom	1959	1959	NãoAtivoEurovisao
5	EurovisionParticipant	…	Spain	1959	2011	AtivoEurovisao
…	…	…	…	…	…	…
184	EurovisionParticipant	…	Ukraine	2022	Ongoing	AtivoEurovisao
185	EurovisionParticipant	…	Russia	2022	Ongoing	AtivoEurovisao

6.2.5 Dimensão Eurovisão

Esta dimensão não apresenta hierarquias.

Tabela 72 - Descrição da Dimensão Eurovisão

Campo	Descrição	Origem dos Dados	Tipo de Dados	Valores	[1]
IDEurovisao	Chave Substituta	ID gerado sequencialmente	C	0-64	-
Numero Edicao	Número da edição da Eurovisão	Tabela Música (campo Value.Eurovision_Number)	T	1-65	-
Total Participante	Total de países participantes na edição	Dado gerado com recurso a linguagem de programação	N	0-43	-
Verificacao VotoJuri	Indica se na edição da Eurovisão houve votos do júri	Dado gerado com recurso a linguagem de programação	C	“Jury Vote” “No Jury Vote”	-
Verificacao Televoto	Indica se na edição da Eurovisao houve votos por televoto	Dado gerado com recurso a linguagem de programação	C	“Televote” “No Televote”	-
Votos PorPais	Indica o número total de pontos que um país atribui	Dado gerado com recurso a linguagem de programação	N	6-116	NULL
PontosMax PorPais	Indica o número máximo de pontos que um país A pode dar a um país B	Dado gerado com recurso a linguagem de programação	N	5-24	NULL
Descricao Pontos	Campo descritivo dos pontos que cada país pode dar	Dado obtido com informacao da wikipedia (*)	T	96-4988	NULL
Max Pontuacao PorPais	Indica a pontuação máxima que uma música pode ter numa edição da eurovisão	Dado gerado com recurso a linguagem de programação	N	48-1032	NULL

Ano	Ano da edição da eurovisão	Tabela Música (campo Value.Eurovision_Number)	D	1956-2021 (1) não há dados de 2020- não se realizou	-
PaisAnfitriao	Pais onde foi realizada a edição da eurovisão	Tabela Música (campo Value.Host_Country)	T	Ex. “Switzerland”	-
CidadeAnfitriap	Cidade onde foi realizada a edição da eurovisão	Tabela Música (campo Value.Host_City)	T	Ex. “Lugano”	-

[1] Valor Inserido quando desconhecido ou não aplicável

Visualização exemplificativa

ID_Euro	(..)	PointsDescription	TotalVotes	Year	HostCountry	HostCity
1	(…)	NULL	NULL	1956	Switzerland	Lugano
2	(…)	10-1	100	1957	West Germany	Frankfurt
3	(…)	10-1	100	1958	Netherlands	Hilversum
4	(…)	10-1	110	1959	France	Cannes
5	(…)	10-1	130	1960	United Kingdom	London
6	(…)	10-1	160	1961	France	Cannes
(…)	(…)	(…)	(…)	(…)	(…)	(…)

6.2.6 Junk Dimension

Na junk dimension não existem hierarquias. Esta tabela foi utilizada para fazer uma verificação simples que vai ser necessária para a primeira tabela de factos, sobre se os países que dão pontos uns aos outros são ou não vizinhos e participaram ou não no mesmo conflito.

Tabela 73 - Descrição da Junk Dimension

Campo	Descrição	Origem dos dados	Tipo de dados	Valores	[1]
IDJunk	Chave Substituta	ID gerado sequencialmente	C	1-9	-
VerificacaoVizinho	Indica se os países em questão são vizinhos	Inserção manual do campo	T	Sim_vizinhos, Nao_vizinhos	NULL
VerificacaoConflito	Indica se os países em questão se encontram em conflito	Inserção manual do campo	T	Sim_conflito, Não_conflito	NULL

[1] Valor Inserido quando desconhecido ou não aplicável

Visualização exemplificativa

IDJunk	VerificacaoVizinho	VerificacaoConflito
1	Sim_vizinhos	Sim_conflito
2	Sim_vizinhos	Nao_conflito
3	Nao_vizinhos	Nao_conflito
4	Nao_vizinhos	Sim_conflito
5	NULL	Sim_conflito
(…)	(…)	(…)

6.3 Medidas Numéricas Aditivas e Não Aditivas

6.3.1 – Medidas numéricas na tabela de factos 1

Grão: No Ano A, o País B deu ao País C o número de Pontos X do Tipo D

Tabela 74 - análise medidas numéricas na tabela de factos 1

Campo	Descrição	Origem dos dados	Intervalos
Pontos	Número de pontos que cada música recebeu	Tabela Eurovisão, campo Points	1- 12

6.3.2 – Medidas numéricas na tabela de factos 2

Grão: Música (País A recebeu o número de Pontos X no Ano B)

Tabela 75 - análise medidas numéricas na tabela de factos 2

Campo	Descrição	Origem dos dados	Intervalos
Pontos	Número de pontos que cada música recebeu	Tabela Musica, campo Sc.	0 - 12
PIB per capita	Valor correspondente ao PIB de cada país por ano	Tabela PIB, campo PIBperCapita	1878.1- 36920.8
CO_2	Número de gases anuais emitidos por cada país.	Tabela Emissões, campo Dados Anuais	0.140 - 1790.34
Turistas	Número total de chegadas de turistas a estabelecimento de alojamento turístico por ano por país.	Tabela Turistas, campo *Number*	56 666 – 140030631
Área	Área total de cada país	Tabela *Area* campo *LandArea*	2.027 – 16389950
População	Número de população residente no país	Tabela População campo Population	13410 – 148538197
Densidade populacional	Densidade populacional referente a cada país	Obtida pela divisão da medida de população pela medida de área	
Emissões por população	Emissões de CO_2 por pessoa em cada país	Obtida pela divisão da medida de emissões pela medida de população	
Emissões por área	Emissões de CO_2 por área	Obtida pela divisão da medida de emissões pela medida de área	
Turismo por área	Turistas por área de um país	Obtida pela divisão da medida de turismo pela medida de área	

Nas tabelas anteriores são referidas diversas medidas podendo estas ser aditivas, ou seja, medidas que fazem sentido somar, medidas semi-aditivas, isto é, que só fazem sentido somar em determinadas dimensões, e medidas não aditivas, cuja soma não apresenta significado.
Como medidas aditivas existem:

- CO_2;
- Turismo.

Como medida semi-aditiva:

- Pontos

Nesta só faz sentido somar os pontos por ano e por país.
Como medidas não aditivas

- PIB per capita
- Área (km^2)
- População
- Densidade Populacional
- Emissões por população
- Emissões por população
- Turismo por área

6.4 Tabela Multivalor

Após ter sido criada a Dimensão Conflito, deparámo-nos com a seguinte questão: e se um país tiver em mais do que um conflito ao mesmo tempo? Para resolver esta interrogação decidimos gerar uma tabela multivalor e uma tabela ponte que nos permitisse, primeiramente agregar todos os conflitos em que um país está envolvido em simultâneo, e seguidamente ligar o grupo de conflitos à tabela de factos.

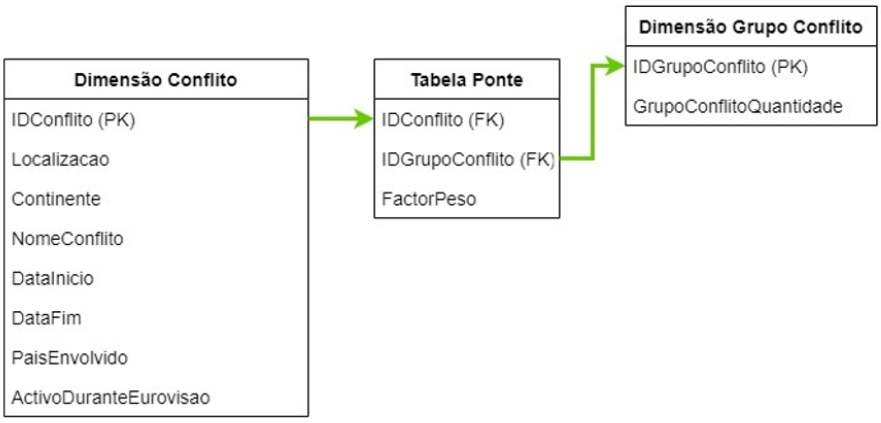

Figura 28 - Tabela Multivalor

6.5 Roleplaying

Roleplaying acontece quando uma dimensão aparece ligada mais do que uma vez à mesma tabela de facto. Aplicando ao nosso projeto, na tabela de factos cujo grão corresponde ao número de pontos, que um país A dá a um país B, num determinado Ano, havendo também a indicação do tipo de pontos que é dado, a dimensão Localização é referida duas vezes, através do país - o país que dá pontos e o país que recebe os pontos.

Como tal, aplicou-se a técnica do roleplaying, existindo apenas uma tabela física - Dimensão Localização - e criou-se duas vistas SQL com atributos específicos para cada caso - Dimensão LocalizaçãoDa e DimensãoLocalizaçãoRecebe.

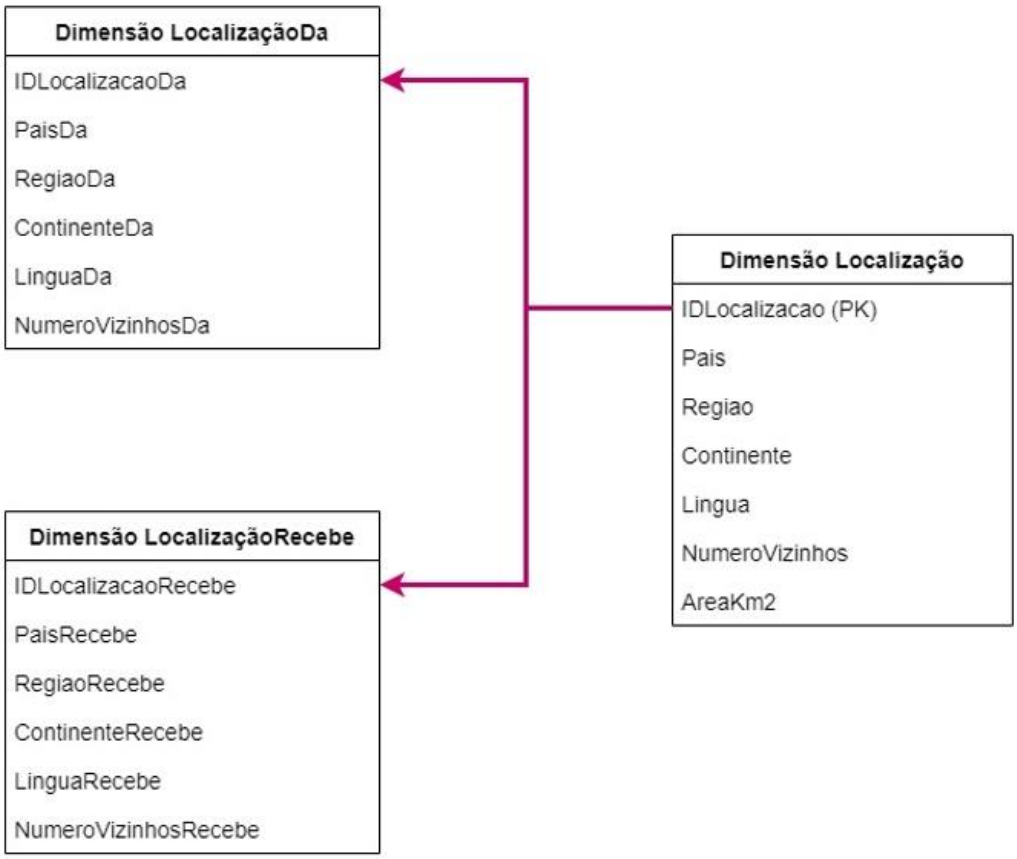

Figura 29: Técnica do role-playing

6.6 Diagrama da Tabela de Factos

Na Figura 30 constata-se uma visão geral do nosso Esquema em Estrela conforme as tabelas de factos e Dimensões criadas. Uma vez que temos duas tabelas de factos, decidimos dividir o esquema separando as tabelas de factos apenas para facilitar a visualização das ligações existentes. Nas ligações, procurámos utilizar uma mesma cor para identificar a relação entre uma dimensão e as tabelas de factos.

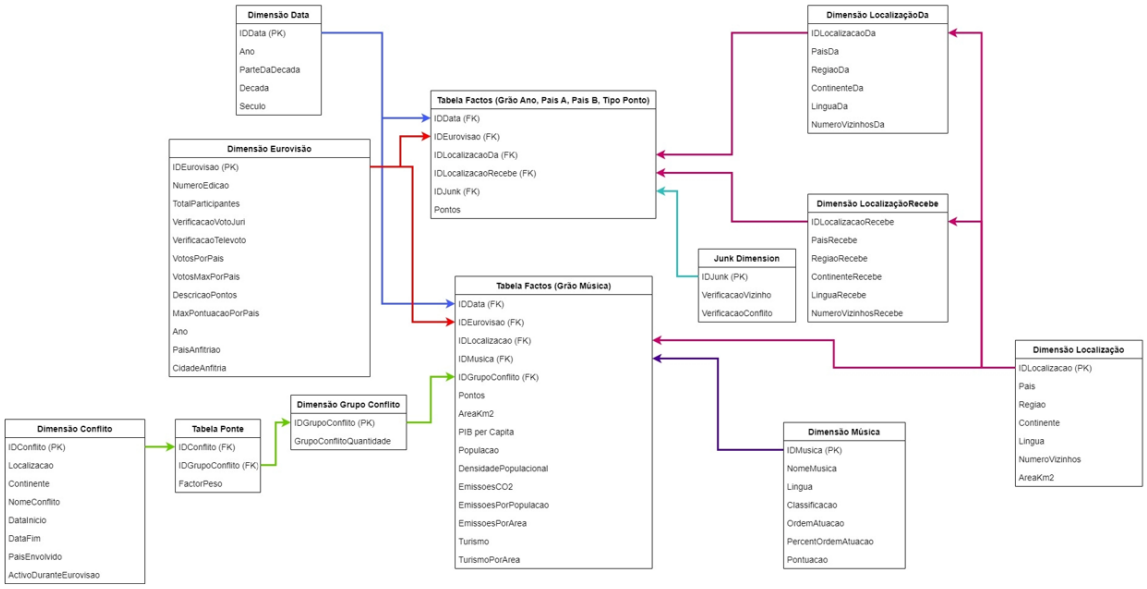

Figura 30- Esquema em estrela global

Na Figura 31 observa-se as ligações entre as dimensões e a tabela de factos com um grão por música

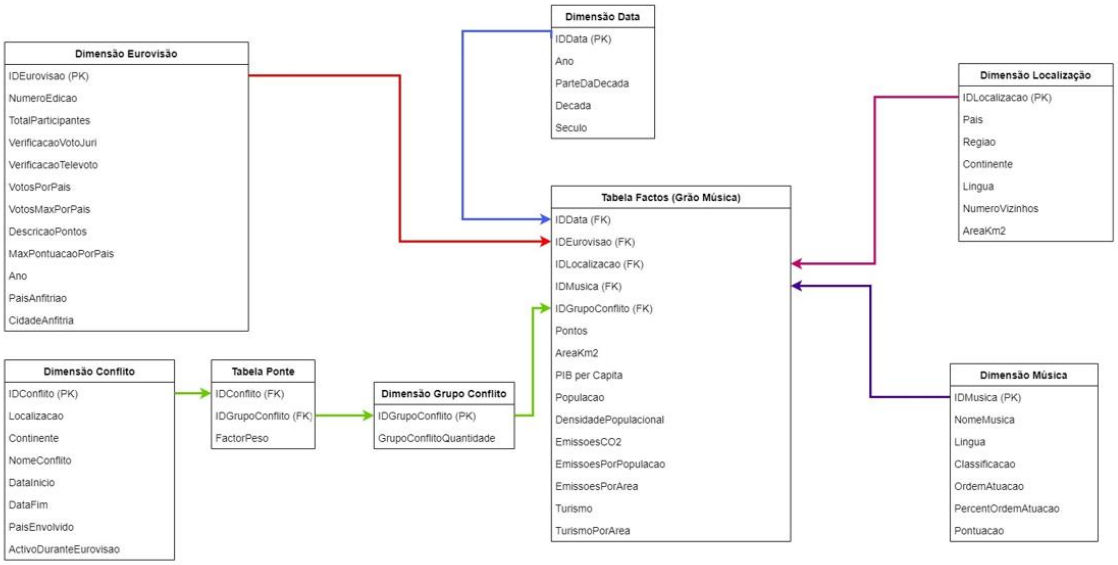

Figura 31 - Tabela de Factos grão Música

Na Figura DD observa-se as ligações entre as dimensões e a tabela de factos com um grão por país, país, tipo, ano.

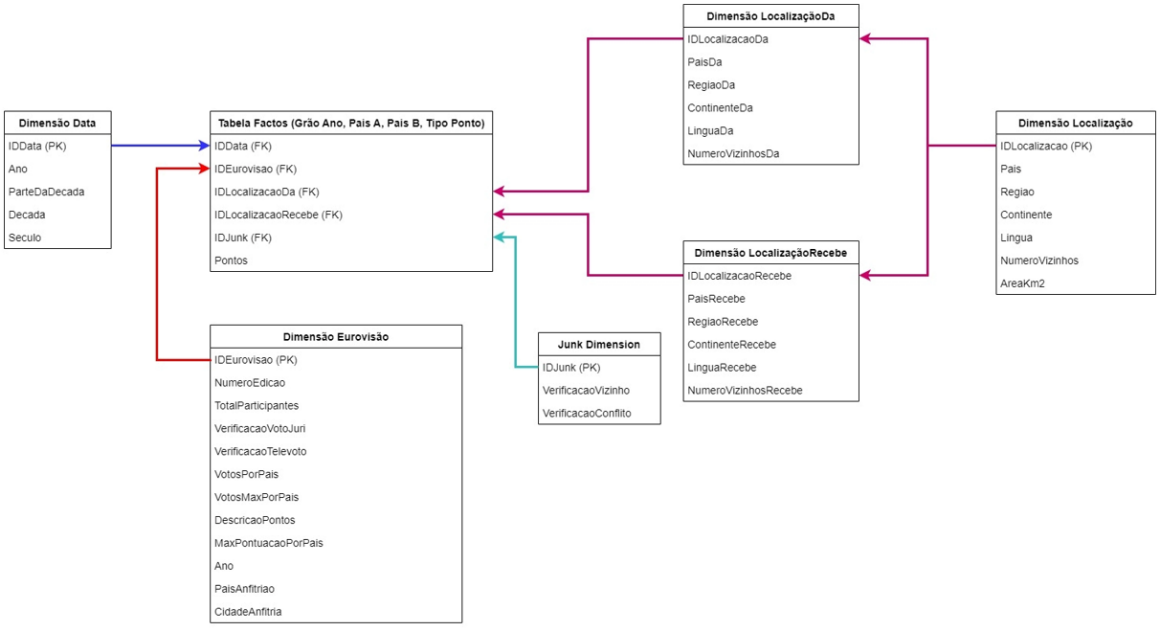

Figura 32 - Tabela de factos grão País, País, Tipo, Ano

Conclusão

Com a elaboração da primeira etapa do projeto desta Unidade Curricular, conseguimos perceber que, por vezes, o formato dos dados disponíveis na internet nem sempre é o mais adequado para o estudo realizado ou que apresenta erros, sendo necessário pequenas alterações. Esta fase permitiu, ainda, uma melhor compreensão dos dados através de análises estatísticas para cada conjunto de dados, como a moda, valor máximo, mínimo, entre outros. Por fim, elaboramos três questões analíticas que pretendemos responder na terceira etapa do projeto através dos dois processos de negócio considerados.

Já na segunda fase, percebemos que foi necessária uma normalização das tabelas, isto é, ter o nome das tabelas numa só língua – a escolhida foi a nossa língua materna – pois estavam em português e em inglês. Eliminamos, ainda, dados das tabelas que não se aplicavam ao projeto, por exemplo o PIB. Na primeira entrega tínhamos 6 tabelas referentes ao PIB, no entanto a que nos interessa é a tabela referida ao PIB per capita. Nesta fase não fizemos referência à tabela dos géneros musicais uma vez que esta apresenta uma reduzida variação pois os géneros mais ouvidos ao longo dos anos eram sempre Pop ou Dance Pop, não apresentando assim, variações que possam ser significativas para o nosso estudo.

Ainda nesta fase, foram definidas 7 dimensões – Data, Localização, Conflitos, Música, Eurovisão, Grupo Conflito e *Junk* - e duas tabelas de factos – cada uma com granularidade diferente. Aqui realizou-se a técnica de *Roleplaying* e dimensões multivalores.

Bibliografia

Discogs. (31 de 03 de 2022). *Discogs*. Obtido de Discogs: https://www.discogs.com/

Eurovision. (31 de 03 de 2022). *Eurovision Events*. Obtido de Eurovision: https://eurovision.tv/events

EurovisionWorld. (31 de 03 de 2022). *Odds Eurovision Song Contest 2022*. Obtido de EurovisionWorld: https://eurovisionworld.com/odds/eurovision

Ferreira, A. (2022). Aulas Teóricas de Integração e Processamento Analítico de Informação.

Kimball, R. (2013). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling.* Wiley.

Reyes, H. (31 de 03 de 2022). *How do Sporting Companies Make Money*. Obtido de BetandBeat: https://betandbeat.com/betting/blog/how-do-betting-companies-make-money/

Wikipedia. (31 de 03 de 2022). *List of Countries and Territories by Land Borders*. Obtido de Wikipedia: https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders

Wikipedia. (31 de 03 de 2022). *List of wars by date*. Obtido de Wikipedia: https://en.wikipedia.org/wiki/Category:Lists_of_wars_by_date