

# PROJETO DE INTEGRAÇÃO E PROCESSAMENTO ANALÍTICO DE INFORMAÇÃO

Etapa 3 – Análise de Dados, Modelação Dimensional,  
Sistema ETL e Relatórios

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

## Conteúdo

Índice de Figuras .....	4
Índice de Tabelas .....	6
Introdução .....	9
1. Fontes de dados .....	10
2. Descrição e análise dos datasets .....	12
2.1 Eurovisão.....	12
2.1.1 Análise estatística .....	12
2.1.2 Erros e dados em falta.....	15
2.2 Top Spotify 2010-2019 .....	15
2.2.1 Análise estatística .....	15
2.2.2 Erros e dados em falta.....	16
2.3 Top Spotify 2020-2021 .....	17
2.3.1 Análise estatística .....	17
2.3.2 Erros e dados em falta.....	18
2.4 Vizinhos .....	18
2.4.1 Análise estatística .....	18
2.4.2 Erros e dados em falta.....	20
2.5 Música .....	20
2.5.1 Análise estatística .....	21
2.5.2 Erros e dados em falta.....	26
2.6 Localização .....	27
2.6.1. Análise estatística.....	27
2.6.2 Erros e dados em falta.....	28
2.7 Turistas.....	29
2.7.1 Análise estatística .....	29
2.7.2 Erros e dados em falta.....	30
2.8 Emissões.....	30
2.8.1 Análise estatística .....	31
2.8.2 Erros e dados em falta.....	33
2.9 PIB .....	33
2.9.1 Análise estatística .....	34
2.9.2 Erros e dados em falta.....	34
2.10 População .....	35
2.10.1 Análise estatística.....	35
2.10.2 Erros e dados em falta.....	36

2.11 Conflitos .....	36
2.11.1 Análise Estatística.....	37
2.11.2 Erros e dados em falta.....	39
2.12 Área.....	39
2.12.1 Análise Estatística.....	39
2.12.2 Erros nos dados.....	40
3. Diagrama Relacional das Fontes de Dados .....	41
4. Processo de Negócio.....	42
5. Questões Analíticas .....	44
6. Modelação Dimensional .....	45
6.1 Declaração do grão e tipo da tabela de factos .....	45
6.1.1 – Dimensões na tabela de factos 1 .....	45
6.1.2 – Dimensões na tabela de factos 2 .....	46
6.2 Tabelas de Dimensão .....	46
6.2.1 Dimensão Localização .....	46
6.2.2 Dimensão Música.....	47
6.2.3 Dimensão Data.....	48
6.2.4 Dimensão Conflitos .....	49
6.2.5 Dimensão Eurovisão.....	50
6.2.6 Junk Dimension.....	52
6.3 Medidas Numéricas Aditivas e Não Aditivas .....	52
6.3.1 – Medidas numéricas na tabela de factos 1 .....	52
6.3.2 – Medidas numéricas na tabela de factos 2 .....	53
6.4 Tabela Multivalue .....	54
6.5 Roleplaying .....	54
6.6 Diagrama da Tabela de Factos.....	55
7. Desenvolvimento dos Programas que Compõe o Sistema ETL .....	57
7.1 Criação das dimensões com recurso a <i>Python</i> .....	57
7.2 Manipulação dos datasets de medidas com recurso a <i>Python</i> .....	61
7.3 Criação da tabela de factos 1.....	62
7.4 Criação da tabela de factos 2 .....	63
7.5 Implementação das tabelas de factos no <i>PostgreSQL</i> e <i>PowerBI</i> .....	64
8. Responsabilidades, Inputs e Outputs dos Programas.....	67
9. Desenho do Diagrama do Sistema ETL .....	68
10. Dimensões e tabela de factos implementados no cubo de dados .....	70
11. Produção de relatórios e resposta às questões analíticas .....	73

11.1 Primeira pergunta analítica – Influência da língua da canção.....	73
11.1.1 Quantidade de países que não se qualificam para a final – em inglês e língua materna....	73
11.1.2 As músicas em inglês obtêm melhores resultados? .....	74
11.1.3 Diferença dos resultados de um país quando canta em inglês versus língua materna 80	
11.2 Segunda pergunta analítica – Influência da demografia e geografia .....	82
11.2.1 Influência do número de vizinhos de um país na quantidade de pontos recebidos ...	82
11.2.2 Entreajuda entre vizinhos ou outros países .....	87
11.2.3 Influência da população vizinha dos países.....	88
11.2.4 Os países com maior PIB têm melhores resultados? .....	88
11.3 Terceira pergunta analítica – Influência das questões da atualidade .....	89
11.3.1 A participação em conflitos diminui a média de pontos que um país recebe? .....	89
11.3.2. Os países “verdes” são mais populares?.....	91
11.3.3 O turismo influencia a votação? .....	91
11.4 Curiosidades .....	92
Conclusão.....	94
Bibliografia .....	95

## Índice de Figuras

Figura 1 - Moda de cidade anfitriã .....	13
Figura 2 - Moda de votos dados.....	14
Figura 3 - Moda de votos recebidos .....	14
Figura 4 - Histograma do número de vizinhos .....	19
Figura 5 - Países com mais participações na Eurovisão a partir do dataset música da Eurovisão.....	21
Figura 6 - Campo de linguagem do dataset música da Eurovisão.....	22
Figura 7 - Músicas com pontuações mais altas no dataset música da Eurovisão .....	23
Figura 8 - Músicas com pontuações mais baixas no dataset música da Eurovisão.....	23
Figura 9 - Países anfitriões do maior número de edições no dataset música da Eurovisão.....	24
Figura 10 - Cidades anfitriãs do maior número de edições no dataset música da Eurovisão .....	24
Figura 11 - Campo de validação da linguagem do dataset música da Eurovisão.....	25
Figura 12 - Distribuição do género musical do dataset música da Eurovisão.....	26
Figura 13 - Distribuição de valores do campo "Continent" do dataset Localização .....	28
Figura 14 - Distribuição de valores do campo "Region" do dataset Localização .....	28
Figura 15 - Histograma número de turistas .....	30
Figura 16 - Histograma do campo "Number" por país .....	31
Figura 17 - Análise do campo "Number" por país.....	32
Figura 18 - Análise do campo “Number” por ano .....	32
Figura 19: Emissões em Portugal por ano .....	33
Figura 20-Histograma do PIB per capita de 1960 a 2020 .....	34
Figura 21 - Crescimento populacional na área Eurovisão .....	35
Figura 22 - Histograma de população nos países da área Eurovisão .....	36
Figura 23 - Localizações com maior ocorrência de conflitos no dataset Conflicts Participants .....	37
Figura 24 - Países com mais participações em conflitos no dataset Conflicts Participants.....	38
Figura 25 - Histograma do campo "LandArea" do dataset Area .....	40
Figura 26 - Diagrama relacional entre tabelas .....	41
Figura 27 - Probabilidade de vencer a Eurovisão em 9/03/2022, 31/03/2022 e 01/05/2022 (EurovisionWorld, 2022).....	43
Figura 28 - Tabela Multivalor .....	54
Figura 29: Técnica do role-playing .....	55
Figura 30- Esquema em estrela global .....	56
Figura 31 - Tabela de Factos grão Música.....	56
Figura 32 - Tabela de factos grão País, País, Tipo, Ano .....	57
Figura 33 - Código Python para Criação da Dimensão Data .....	58
Figura 34 - Código Python para Criação das Dimensões Junk.....	59
Figura 35 - Código Python para Criação da Dimensão Localização.....	59
Figura 36 - Código Python para Criação da Dimensão Conflitos .....	60
Figura 37 - Código Python para Criação da Dimensão Música .....	60
Figura 38 - Formato inicial do dataset .....	61
Figura 39 - Código Python para Manipulação dos Datasets de Medidas .....	61
Figura 40 - Formato resultante do código Python aplicado ao Dataset .....	62
Figura 41 – Excerto do Código Python para Criação da Tabela de Factos 1.....	63
Figura 42 - Excerto do Código Python para Criação da Tabela de Factos 2 .....	64
Figura 43 – Excerto dos comandos para criação das dimensões no PostgreSQL .....	65
Figura 44 - Criação das tabelas de factos no PostgreSQL.....	65
Figura 45 - Criação das vistas materializadas no PostgreSQL.....	66

Figura 46 - Vistas implementadas no PostgreSQL .....	66
Figura 47 - Diagrama dos programas do sistema ETL.....	67
Figura 48 - Diagrama do sistema ETL.....	69
Figura 49: Implementação das dimensões e tabelas de factos utilizando o Power Bi.....	70
Figura 50: Implementação das dimensões e tabelas de factos utilizando o postgreSQL .....	71
Figura 51: Tabela de factos 1 .....	71
Figura 52: Tabela de factos 2 .....	71
Figura 53 - Comando SQL para a criação de um cubo.....	72
Figura 54 - Exemplo de resultados do cubo .....	72
Figura 55: Código SQL utilizado e respetivo resultado obtido para a língua Não Inglês.....	73
Figura 56: Código SQL utilizado e respetivo resultado obtido para a língua inglês .....	74
Figura 57: Código SQL utilizado e respetivo resultado obtido para a língua Mixed .....	74
Figura 58: Média das pontuações por tipo de linguagem da música .....	75
Figura 59: Média de pontos quando não existe regra de linguagem .....	75
Figura 60: Média de pontos quando existe regra de linguagem.....	75
Figura 61: Classificação média por língua quando existe regra de língua materna .....	76
Figura 62: Classificação média por língua quando não existe regra de língua materna .....	76
Figura 63: Média das classificações por década .....	77
Figura 64 - Primeiros classificados por língua ao longo das décadas .....	78
Figura 65 - Top 5 classificados por língua ao longo das décadas .....	79
Figura 66: Código SQL utilizado e respetivos resultados obtidos .....	80
Figura 67: Comparação das médias das classificações obtidas por cada país por língua .....	80
Figura 68: Comparação de resultados obtidos pelo mesmo país cantando músicas em inglês e não inglês. ....	81
Figura 69: Comparação de resultados obtidos pelo mesmo país cantando músicas em inglês e mistura do inglês com a sua língua materna.....	81
Figura 70: Número de vizinhos que cada país tem e a respetiva média de pontos que obtém no festival. ....	82
Figura 71: Soma da pontuação total que cada país dá ao seu vizinho. ....	83
Figura 72: Soma de pontos divida por número de países com número x de vizinhos versus número de vizinhos .....	83
Figura 73: Classificação média dos países de acordo com o seu número de vizinhos. ....	84
Figura 74: Soma da Classificação por número de vizinhos. ....	84
Figura 75: Classificação por número de vizinhos dividido pelo número de países com x vizinhos....	84
Figura 76 - Comandos SQL para procura do vizinho mais valioso .....	85
Figura 77 - Resultado obtido .....	85
Figura 78: Código SQL utilizado para procura do país que mais pontos recebe dos seus vizinhos... ...	85
Figura 79: Mapa de pontos dados aos países vizinhos em todas as edições da Eurovisão .....	86
Figura 80 - Média de pontos que cada país recebe por edição pelos seus vizinhos.....	86
Figura 81: Comparação das médias das classificações obtidas por cada país com a população de cada país. ....	88
Figura 82: Comparação das médias das classificações obtidas por cada país com o PIBPerCapita de cada país. ....	89
Figura 83: Código SQL utilizado e respetivos resultados obtidos .....	89
Figura 84: Código SQL utilizado e respetivos resultados obtidos .....	90
Figura 85: Comparação das médias das classificações obtidas por cada país tendo em conta se está perante um conflito ou não. ....	90

Figura 86: Comparação das médias das classificações obtidas por cada país com as emissões de CO <sub>2</sub> de cada país. ....	91
Figura 87: Comparação das médias das classificações obtidas por cada país com o turismo por área. ....	91
Figura 88 - Classificação média no concurso por ordem de atuação .....	92
Figura 89 - Classificação média no concurso por ordem de atuação na década de 2000.....	93
Figura 90 -Classificação média no concurso por ordem de atuação na década de 2010.....	93
Figura 91 - Classificação média no concurso por ordem de atuação na década de 2020 .....	93

## Índice de Tabelas

Tabela 1- Conjunto de dados e a sua descrição .....	10
Tabela 2 - Descrição dos campos do dataset da Eurovisão .....	12
Tabela 3 - Descrição do campo "Index" do dataset da Eurovisão.....	12
Tabela 4 - Descrição do campo "Host City" do dataset da Eurovisão .....	12
Tabela 5 - Descrição do campo "Year" do dataset da Eurovisão .....	13
Tabela 6 - Descrição do campo "Points Type" do dataset da Eurovisão .....	13
Tabela 7 - Descrição do campo "From" do dataset da Eurovisão .....	13
Tabela 8 - Descrição do campo "To" do dataset da Eurovisão .....	14
Tabela 9 - Descrição do campo "Points" do dataset da Eurovisão.....	14
Tabela 10 - Descrição dos campos do dataset Spotify 2010-2019.....	15
Tabela 11 - Descrição do campo "Index" do dataset Spotify 2010-2019 .....	15
Tabela 12 - Descrição do campo "Artist" do dataset Spotify 2010-2019 .....	15
Tabela 13 - Descrição do campo "Top Genre" do dataset Spotify 2010-2019 .....	16
Tabela 14 - Descrição do campo "Year" do dataset Spotify 2010-2019 .....	16
Tabela 15 - Descrição do campo "pop" do dataset Spotify 2010-2019.....	16
Tabela 16 - Moda e número de ocorrências para o género musical mais ouvido entre os anos 2010 e 2019.....	16
Tabela 17 - Descrição dos campos do dataset Spotify 2020-2021.....	17
Tabela 18 - Descrição do campo "Index" do dataset Spotify 2020-2021 .....	17
Tabela 19 - Descrição do campo "Artist" do dataset Spotify 2020-2021 .....	17
Tabela 20 -Moda e número de ocorrências para o género musical mais ouvido em 2021.....	18
Tabela 21 - Descrição do campo "Popularity" do dataset Spotify 2020-2021.....	18
Tabela 22 - Descrição dos campos do dataset Vizinhos .....	18
Tabela 23 - Descrição do campo "ID" do dataset Vizinhos.....	19
Tabela 24 - Descrição do campo "No. of Neighbours" do dataset Vizinhos .....	19
Tabela 25 - Descrição do campo "Neighbours" do dataset Vizinhos .....	19
Tabela 26 - Descrição dos campos do dataset Música .....	20
Tabela 27 - Descrição dos campos adicionados ao dataset Música.....	20
Tabela 28 - Descrição do campo "ID" do dataset Música.....	21
Tabela 29 - Descrição do campo "Country" do dataset Música.....	21
Tabela 30 - Descrição do campo "Participation" do dataset Música .....	21
Tabela 31 - Descrição do campo "Artist" do dataset Música.....	22

Tabela 32 - Descrição do campo "Language" do dataset Música .....	22
Tabela 33 - Descrição do campo "PI" do dataset Música .....	22
Tabela 34 - Descrição do campo "Sc" do dataset Música.....	23
Tabela 35 - Descrição do campo "Eurovision_Number" do dataset Música .....	23
Tabela 36 - Descrição do campo "Year" do dataset Música .....	23
Tabela 37 - Descrição dos campos "Host Country" e "Host City" do dataset Música .....	24
Tabela 38 - Descrição do campo "EnglishNonEnglish" do dataset Música.....	25
Tabela 39 - Descrição do campo "Running Order" do dataset Música .....	25
Tabela 40 - Descrição do campo "Genre" do dataset Música .....	26
Tabela 41 - Descrição dos campos do dataset Localização .....	27
Tabela 42 - Descrição do campo "ID" do dataset Localização .....	27
Tabela 43 - Descrição dos campos do dataset Turistas .....	29
Tabela 44- Descrição do campo "Geo (labels)" do dataset Turistas .....	29
Tabela 45 - Descrição do campo "Time" do dataset Turistas .....	29
Tabela 46 - Descrição do campo "Number" do dataset Tourists .....	29
Tabela 47 - Descrição dos campos do dataset Emissões .....	30
Tabela 48 - Descrição do campo "Year" do dataset Emissoes .....	31
Tabela 49 - Descrição do campo "Dados Anuais" do dataset Emissoes .....	31
Tabela 50 - Descrição dos campos do dataset PIB .....	33
Tabela 51-Descrição do campo "Year" do dataset PIB .....	34
Tabela 52 - Descrição do campo "PIBpercapita" do dataset PIB .....	34
Tabela 53 - Descrição dos campos do dataset Populacao .....	35
Tabela 54 - Descrição do campo "Year" do dataset Populacao .....	35
Tabela 55 - Descrição do campo "População" do dataset Populacao.....	35
Tabela 56 - Descrição dos campos do dataset Conflitos .....	36
Tabela 57 - Descrição dos novos campos da tabela do dataset Conflitos.....	36
Tabela 58 - Descrição do campo "ID" do dataset Conflitos .....	37
Tabela 59 - Descrição do campo "Conflict Location" do dataset Conflitos .....	37
Tabela 60 - Descrição do campo "EurovisionCountry" do dataset Conflitos.....	38
Tabela 61 - Descrição dos campos "Start Date" e "End Date" do dataset Conflitos .....	38
Tabela 62 - Descrição do campo "Participant" do dataset Conflitos .....	38
Tabela 63 - Descrição dos campos do dataset Área.....	39
Tabela 64 - Descrição do campo "Year" do dataset Area .....	40
Tabela 65 - Descrição do campo "LandArea" do dataset Area .....	40
Tabela 66 - Descrição das dimensões da tabela de factos 1.....	45
Tabela 67 - Descrição das dimensões da tabela de factos 2.....	46
Tabela 68 - Descrição da Dimensão Localização .....	46
Tabela 69 - Descrição da Dimensão Música.....	47
Tabela 70 - Descrição da Dimensão Data .....	48
Tabela 71 - Descrição da Dimensão Conflitos .....	49
Tabela 72 - Descrição da Dimensão Eurovisão .....	50
Tabela 73 - Descrição da Junk Dimension.....	52
Tabela 74 - análise medidas numéricas na tabela de factos 1.....	52
Tabela 75 - análise medidas numéricas na tabela de factos 2 .....	53
Tabela 76- Combinações mais populares de países .....	87

### *Principais alterações realizadas – 1<sup>a</sup> para 2<sup>a</sup> fase de entrega*

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).

### *Principais alterações realizadas – 2<sup>a</sup> para 3<sup>a</sup> fase de entrega*

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

- Inicialmente, depois de obtermos os datasets com relevância para o nosso tema, começámos a trabalhar os dados em Excel. Foi neste programa que corrigimos erros, que realizámos análises estatísticas e que construímos as nossas dimensões, porém, ao falarmos com o docente da cadeira apercebemo-nos que a criação das Dimensões e Tabelas de Factos deveria ser feita de uma maneira mais automática, como tal, voltámos a utilizar os datasets iniciais e com recurso à linguagem Python construímos novamente as dimensões e Tabelas de Factos passando tudo para código. No entanto, mantivemos toda a análise, correção e criação de campos que tinha sido feita nos datasets originais em Excel de forma a poupar tempo e evitar repetir o trabalho já feito.
- Nesta fase do trabalho decidimos descartar os dados relativos ao género de música, uma vez que estes tinham muita informação em falta, o que não permitiu responder às nossas perguntas analíticas.
- Analisando o diagrama da Tabela de Factos e os dados que temos relativos aos conflitos entendemos que a melhor abordagem a se ter seria simplificar a dimensão Conflitos no esquema da Tabela de Factos 2 (grão: música) visto que não precisamos de tanto detalhe para as interrogações. Já no esquema da Tabela de Factos 1 eliminámos a verificação dos conflitos por não haverem dados suficiente, dado que muitos poucos países se encontravam no mesmo conflito durante a sua participação na eurovisão.

## 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.

A última fase do projeto foi a criação e implementação do sistema ETL com recurso à linguagem de programação python para trabalhar os ficheiros de entrada do sistema e com utilização do software PostgreSQL, PowerBI, ArcMap e Excel para gerar relatórios analíticos. Depois de implementado o sistema ETL foi feita uma análise sobre as várias perguntas que tinham sido colocadas no início do projeto com o âmbito de obter respostas para as relações propostas.

## 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
<b>Eurovision finals voting results 1957-2021</b> “Eurovisao.xlsx”	Dataset retirado do Kaggle com os votos de todos os países na Eurovisão de 1957 a 2021	.csv	Download direto	<a href="https://www.kaggle.com/orianao/eurovision-finals-voting-results-19572021">https://www.kaggle.com/orianao/eurovision-finals-voting-results-19572021</a>
<b>Top Spotify (2010-2019)</b> “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	<a href="https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year">https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year</a>
<b>Top Spotify (2020-2021)</b> “Genero.xlsx”	Dataset retirado do Kaggle com as músicas mais ouvidas do Spotify nos anos 2020 e 2021.	.csv	Download direto	<a href="https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021">https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021</a>
<b>Land Borders (Neighbours)</b> “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	<a href="https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders">https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders</a>
<b>Eurovision song lyrics 1956-2021</b> “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	<a href="https://www.kaggle.com/minitree/eurovision-song-lyrics">https://www.kaggle.com/minitree/eurovision-song-lyrics</a>
<b>Countries and Territories</b> “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	<a href="https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/table/">https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/table/</a>
<b>Tourist</b> “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	<a href="https://ec.europa.eu/eurostat/databrowser/view/tour_occ_ar/nat/default/table?lang=en">https://ec.europa.eu/eurostat/databrowser/view/tour_occ_ar/nat/default/table?lang=en</a>

<b>Historical emissions</b> “Emissões.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	<a href="https://www.kaggle.com/datasets/ankanhore545/carbon-dioxide-emissions-of-the-world">https://www.kaggle.com/datasets/ankanhore545/carbon-dioxide-emissions-of-the-world</a>
<b>World GDP</b> <b>(PIB 1960-2020)</b> “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	<a href="https://www.kaggle.com/datasets/zgrcemta/world-gdp-gdp-per-capita-and-annual-growths">https://www.kaggle.com/datasets/zgrcemta/world-gdp-gdp-per-capita-and-annual-growths</a>
<b>World Population</b> <b>(1960-2020)</b> “Populacao.xlsx”	Dataset retirado do <i>The World Bank</i> com informação sobre a população de todos os países e regiões do mundo, de 1960 até 2020.	.xls	Download direto	<a href="https://databank.worldbank.org/reports.aspx?source=2&amp;series=SP.POP.TOTL&amp;country=WLD#">https://databank.worldbank.org/reports.aspx?source=2&amp;series=SP.POP.TOTL&amp;country=WLD#</a>
<b>Conflicts Participants</b> “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	<a href="https://www.kaggle.com/datasets/guybarrash/war-conflicts-and-nations-who-took-part-in-them?select=Conflicts+participants.csv">https://www.kaggle.com/datasets/guybarrash/war-conflicts-and-nations-who-took-part-in-them?select=Conflicts+participants.csv</a>
<b>Land Area</b> “Area.xlsx”	Dataset retirado do <i>The World Bank</i> com informação sobre as áreas de todos os países e regiões do mundo, de 1961 até 2021	.xls	Download direto	<a href="https://data.worldbank.org/indicator/A.G.LND.TOTL.K2">https://data.worldbank.org/indicator/A.G.LND.TOTL.K2</a>

## 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 “Eurovision.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.

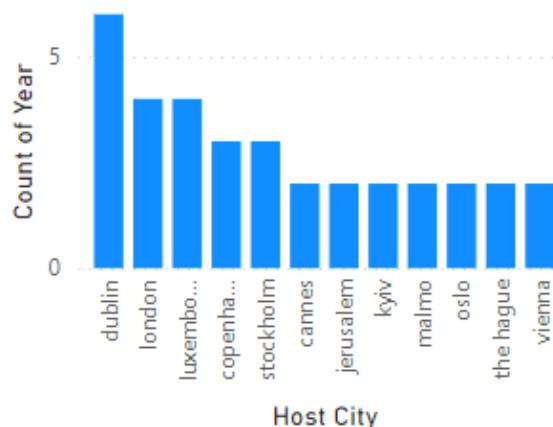
##### 2.1.1.2 – Host City

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.

**Count of Year by Host City**



*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 França, estando em falta nesta lista a Itália que não participou durante um longo período nos anos 90.

### Count of From by From

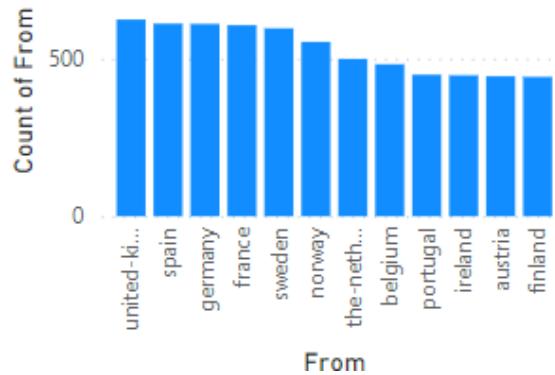


Figura 2 - Moda de votos dados

#### 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.

### Count of To by To

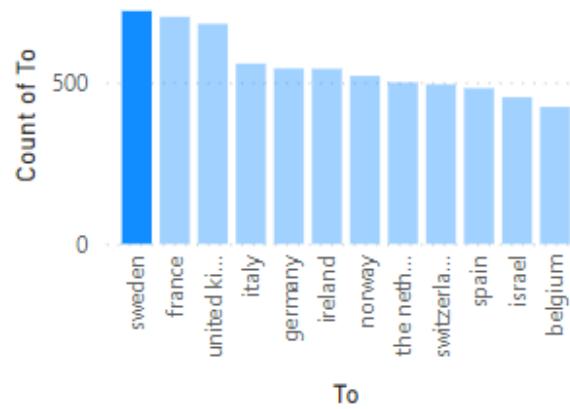


Figura 3 - Moda de votos recebidos

#### 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"	11
	"Maroon 5"	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	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”.

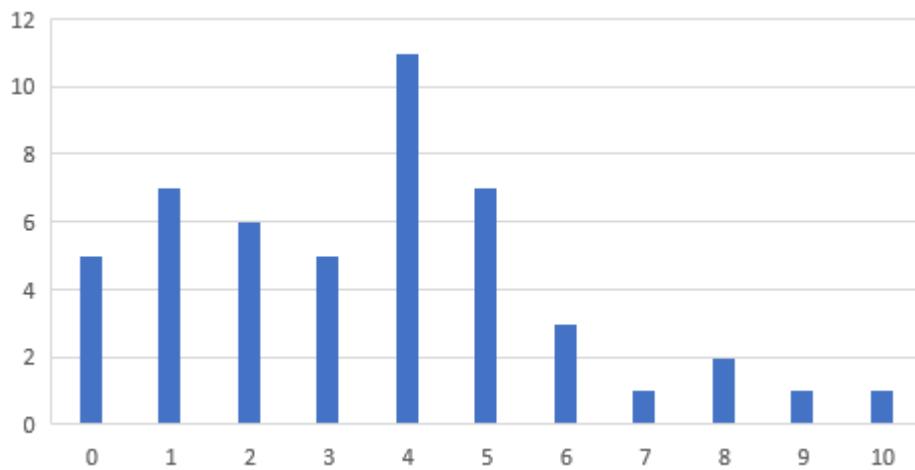


Figura 4 - Histograma do número de vizinhos

#### 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
<b>1</b>	ID	Categórico	Identificador único	176
<b>2</b>	Country	Texto	Nome do País	Portugal
<b>3</b>	Participation	Categórico	Número cumulativo de participações do país	5
<b>4</b>	Artist	Texto	Nome do artista	Carlos Mendes
<b>5</b>	Song	Texto	Nome da música	Verão
<b>6</b>	Language	Texto	Linguagem da música	Portuguese
<b>7</b>	Pl.	Categórico	Classificação Final	11
<b>8</b>	Sc.	Número	Pontuação final	5
<b>9</b>	Eurovision_Number	Categórico	Número de edição	13
<b>10</b>	Year	Número	Ano da edição	1968
<b>11</b>	Host_Country	Texto	Pais organizador	United Kingdom
<b>11</b>	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
<b>12</b>	EnglishNonEnglish	Categórico	Indica se a linguagem da música é inglês, não inglês ou mistura.	English
<b>13</b>	RunningOrder	Número	Número na ordem do concurso em que a música tocou.	1
<b>13</b>	ROpercent	Categórico	Razão entre a ordem dentro do concurso em que a música tocou e o número total de músicas na final.	0.64
<b>15</b>	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

Count of Value.Country by Value.Country

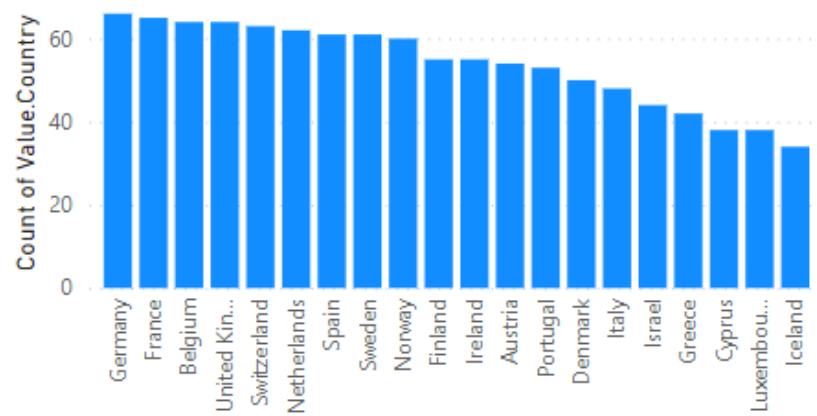


Figura 5 - Países com mais participações na Eurovisão a partir do dataset música da Eurovisão

### 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 excepcionalmente 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

Count of Value.Language by Value.Language

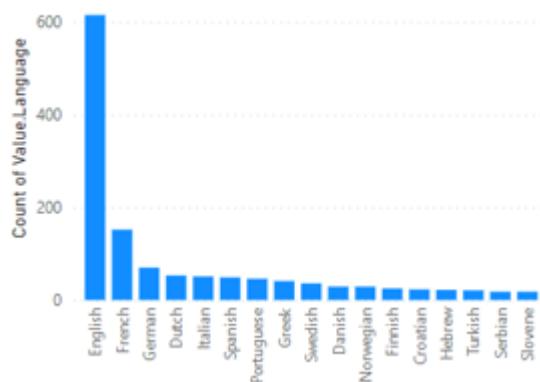


Figura 6 - Campo de linguagem do dataset música da Eurovisão

#### 2.5.1.7 – PI.

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 "PI" do dataset Música

Campo	Média	Máximo	Mínimo	Desvio Padrão
PI	11.26	26	1	6.68

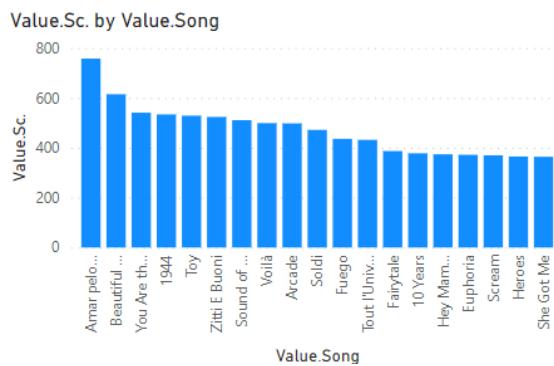
#### 2.5.1.8 – Sc.

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].

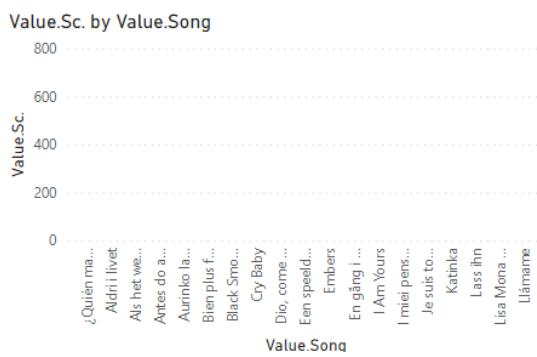
O valor máximo de pontos foi atingido em 2017 com a música “Amar pelos Dois”, interpretada por Salvador Sobral por Portugal. No lado oposto do espectro são várias as músicas e os países que obtiveram 0 pontos no concurso. A pontuação media é de 71 pontos, existindo um elevado desvio padrão.

*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
<b>Host Country</b>	United Kingdom	8
<b>Host City</b>	Dublin	6

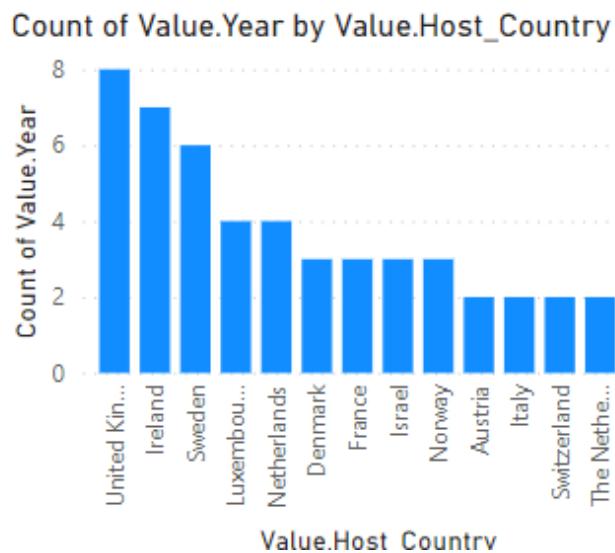


Figura 9 - Países anfitriões do maior número de edições no dataset música da Eurovisão

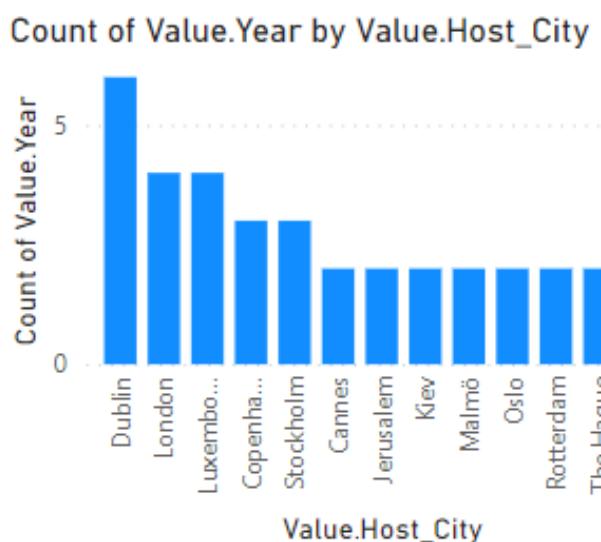


Figura 10 - Cidades anfitriãs do maior número de edições no dataset música da Eurovisão

#### 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

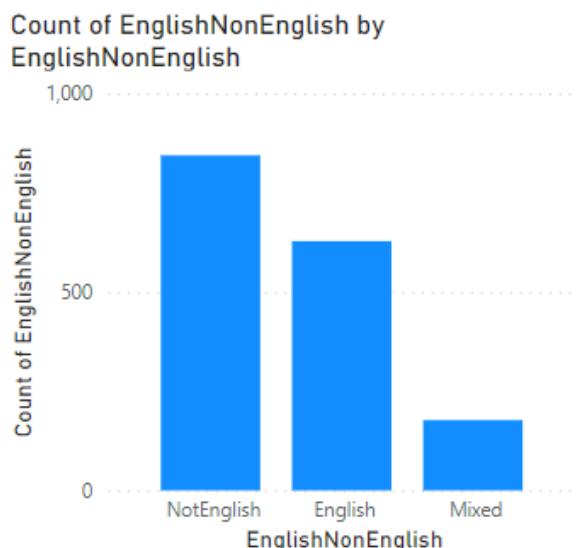


Figura 11 - Campo de validação da linguagem do dataset música da Eurovisão

#### 2.5.1.13 – RunningOrder e ROpercent

Da análise 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 análise 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	# Occurrences
Genre	Pop	122

Count of Genre by Genre

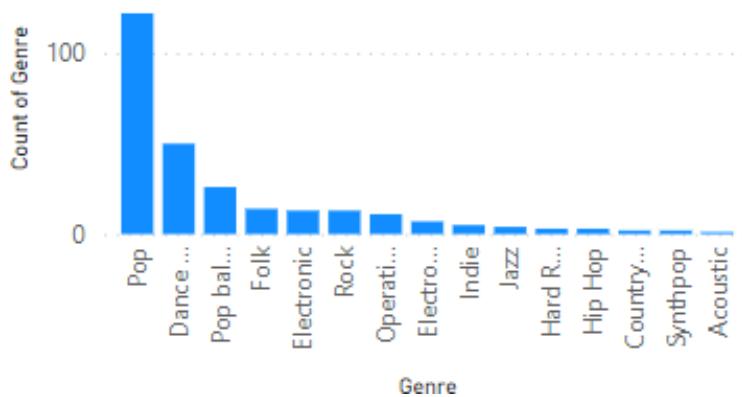


Figura 12 - Distribuição do género musical do dataset música da Eurovisão

#### 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 pretendemos 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 África. Existe ainda o caso da Austrália, que foi convidada a participar em 2015 e desde então tem participado todos os anos.

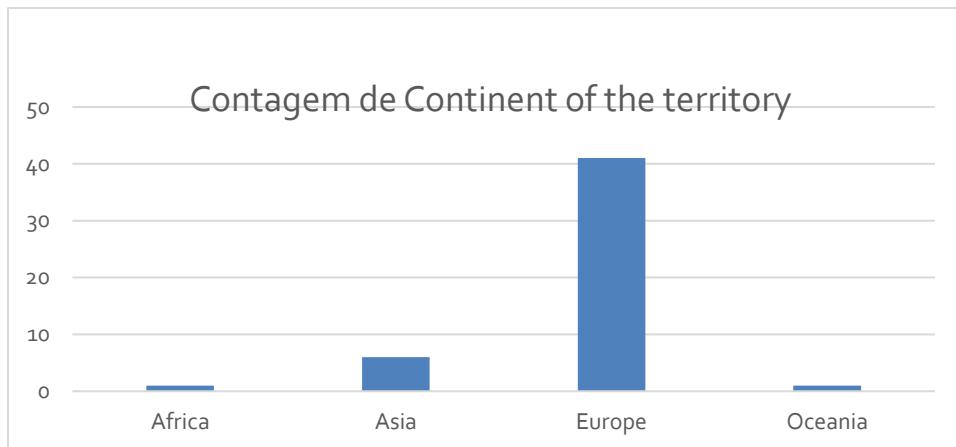


Figura 13 - Distribuição de valores do campo "Continent" do dataset Localização

#### 2.6.1.4 – Region

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

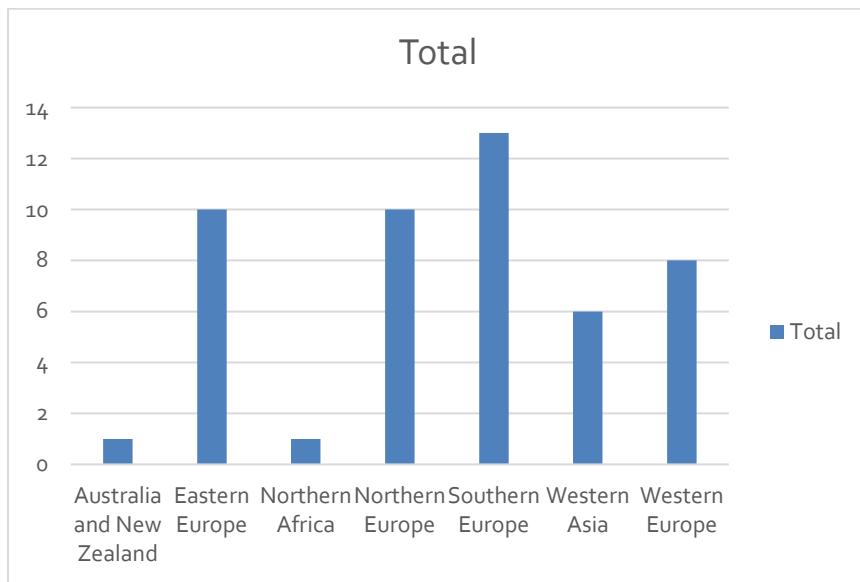


Figura 14 - Distribuição de valores do campo "Region" do dataset Localização

## 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

Analizando 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

Analizando 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

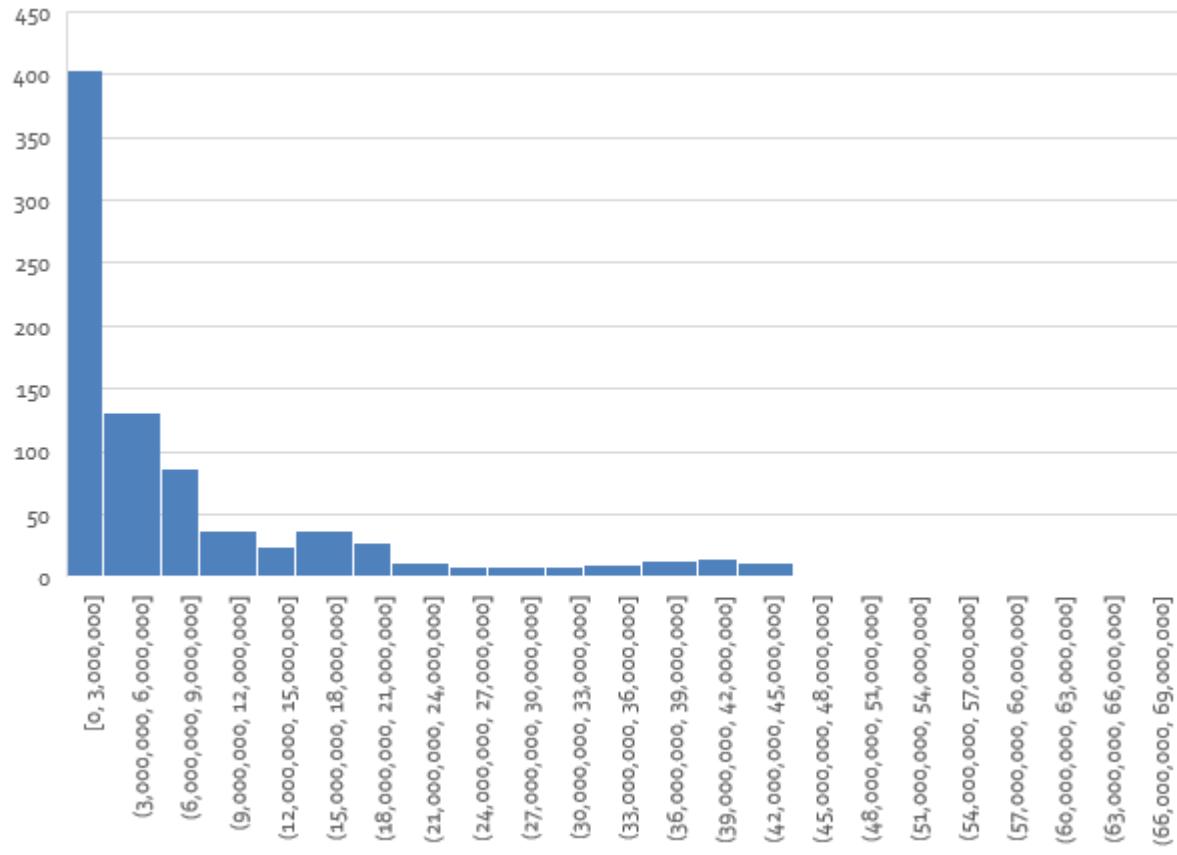


Figura 15 - Histograma número de turistas

### 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 “Emissões.xlsx” guarda os dados de emissão de CO<sub>2</sub> de cada país no mundo. Este documento é constituído pelos seguintes campos.

Tabela 47 - Descrição dos campos do dataset Emissões

#	Campo	Texto	Descrição	Exemplo
<b>1</b>	Country	Texto	Nome do país	Italy
<b>2</b>	Gas	Texto	Tipo de gás emitido	CO2
<b>2</b>	Unit	Texto	Unidade de medida	MtCO <sub>2</sub> e
<b>3</b>	Year	Número	Ano a que se refere a emissão	2000
<b>4</b>	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<sub>2</sub>” e “MtCO<sub>2</sub>”.

### 2.8.1.3 – Year

Analizando o campo “Year”, conseguimos perceber o intervalo de tempo do dataset, visível na tabela seguinte.

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

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 Emissões

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

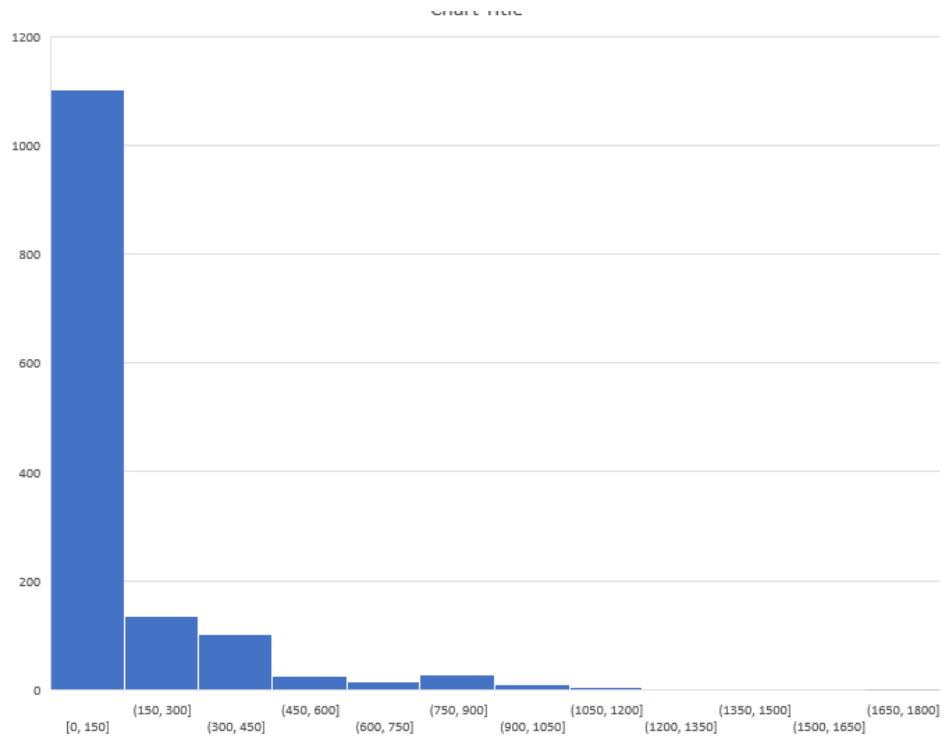


Figura 16 - Histograma do campo "Number" por país

Ao observar a figura 16, é visível que a maioria dos países emite valores de CO<sub>2</sub> baixos existindo alguns outliers.

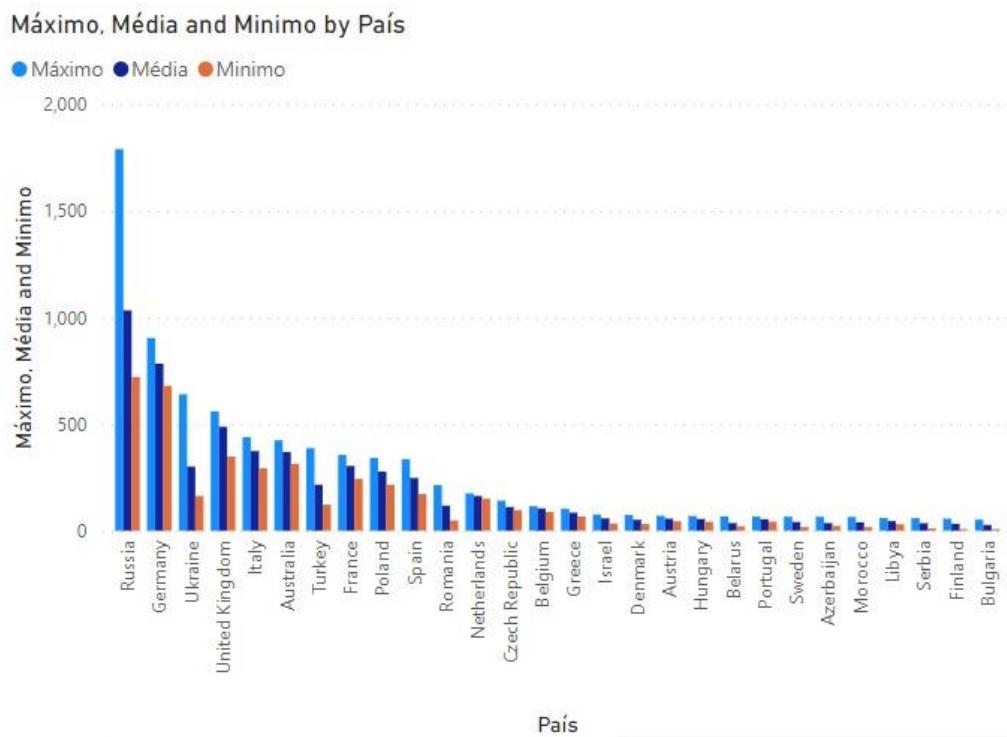


Figura 17 - Análise do campo "Number" por país

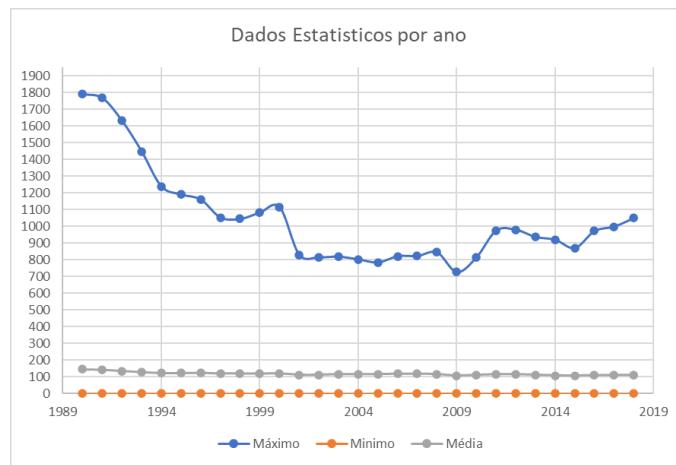


Figura 18 - Análise do campo “Number” por ano

Ao observar a figura “Dados estatísticos por ano” conseguimos determinar que o ano em que houve mais emissões de CO<sub>2</sub> 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<sub>2</sub> 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<sub>2</sub> 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<sub>2</sub> em Portugal ao longo do tempo e um histograma das emissões de CO<sub>2</sub> de todos os países. Tal é possível observar nos dois gráficos seguintes.

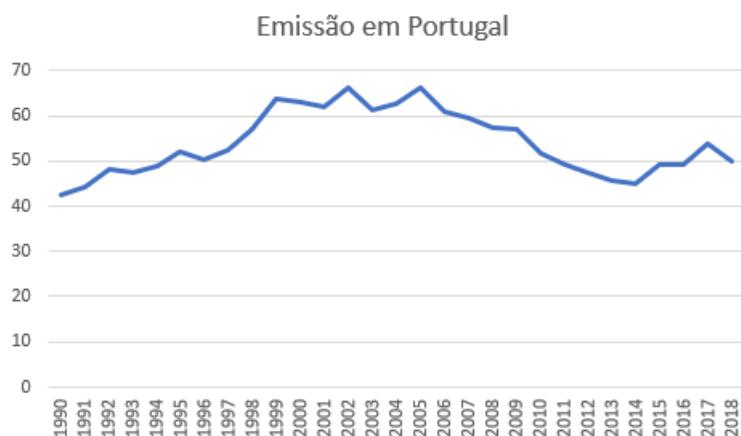


Figura 19: Emissões em Portugal por ano

### 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<sub>2</sub> 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

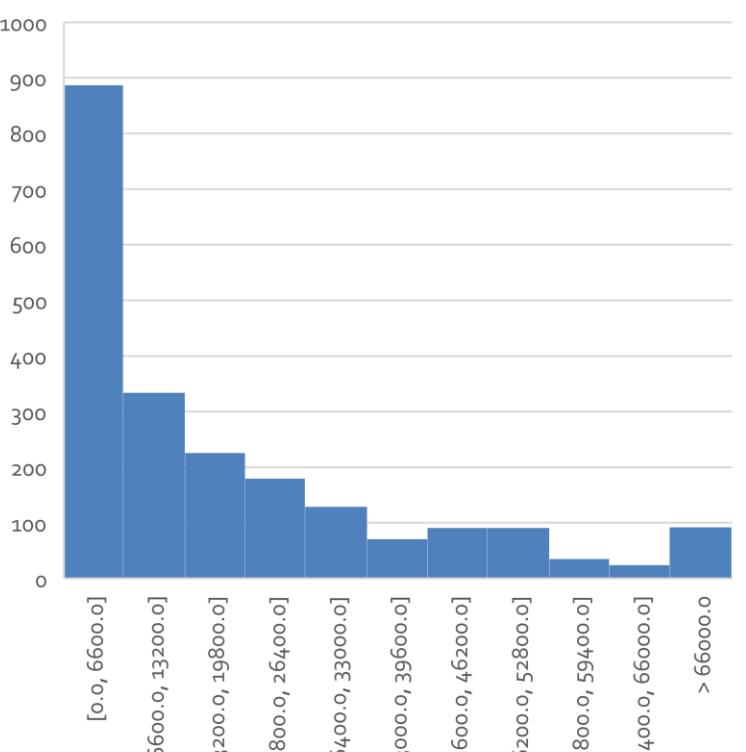


Figura 20-Histograma do PIB per capita de 1960 a 2020

## 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” contém 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.

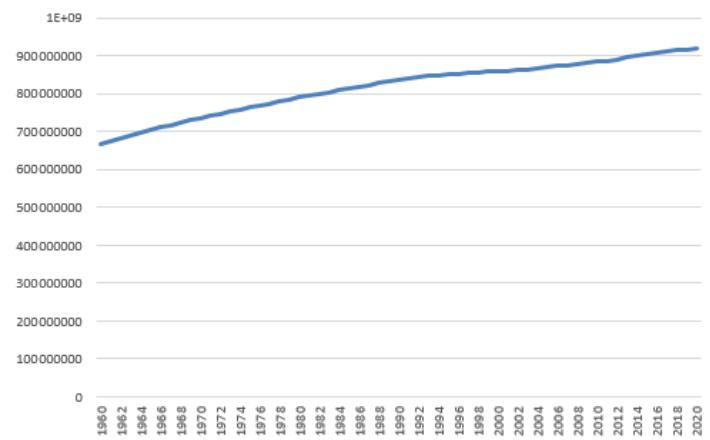


Figura 21 - Crescimento populacional na área Eurovisão

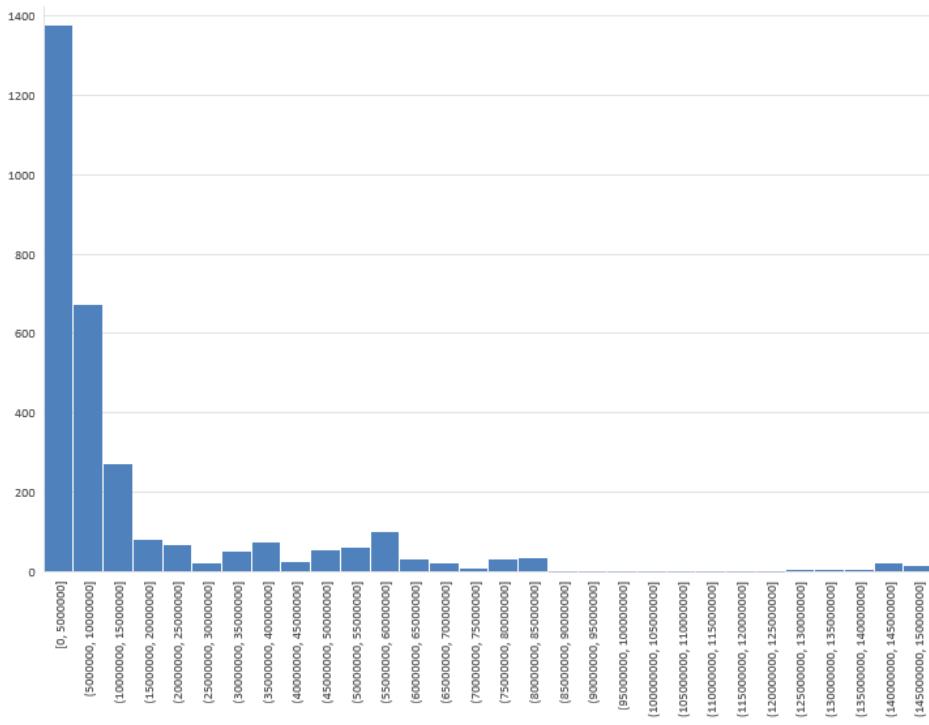


Figura 22 - Histograma de população nos países da área Eurovisão

### 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
<b>Conflict</b>	Texto	Nome do conflito	Gulf War
<b>Start Date</b>	Data	Data de início	1990-08-02T00:00:00Z
<b>End Date</b>	Data	Data de fim	1996-10-24T00:00:00Z
<b>Participant1</b> (...) <b>Participant20</b>	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
<b>1</b>	ID	Categórico	Identificador único	1
<b>2</b>	Conflict Location	Texto	País ou local onde o conflito se realizou maioritariamente	Morocco

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

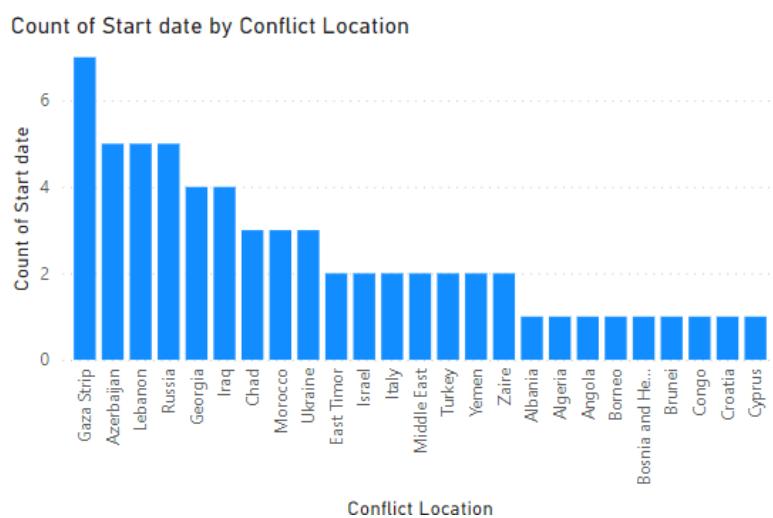


Figura 23 - Localizações com maior ocorrência de conflitos no dataset Conflicts Participants

#### 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.

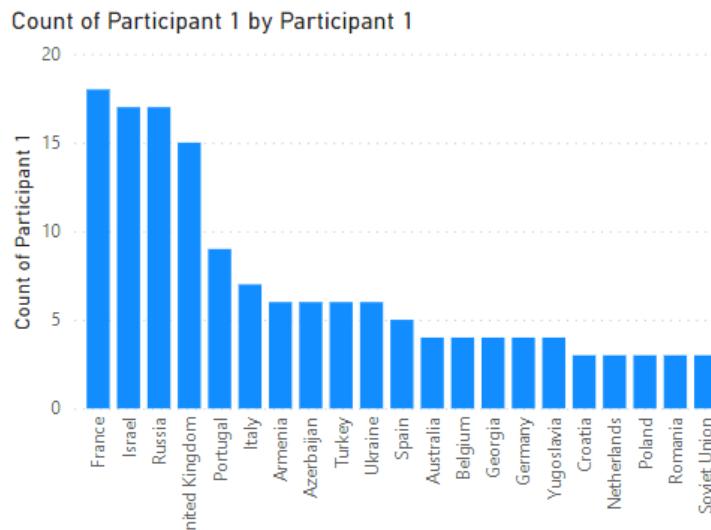


Figura 24 - Países com mais participações em conflitos no dataset Conflicts Participants

### 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 compararam 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 (km <sup>2</sup> )	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

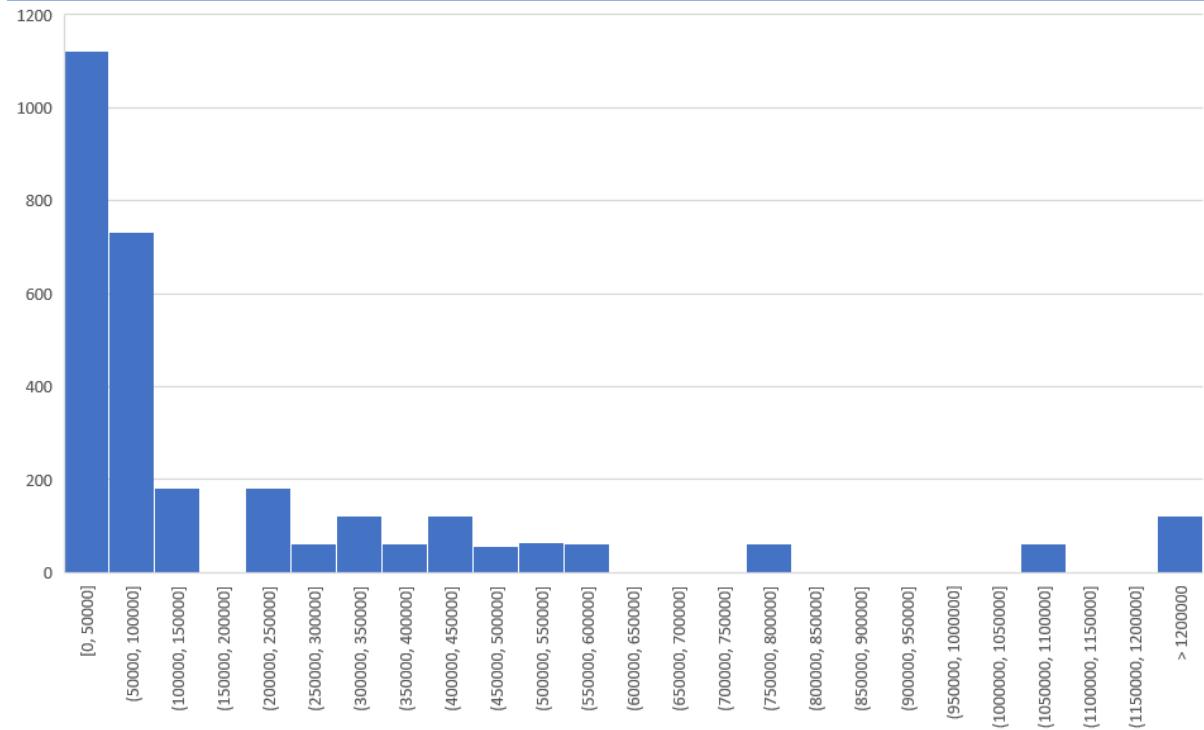


Figura 25 - Histograma do campo "LandArea" do dataset Area

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.

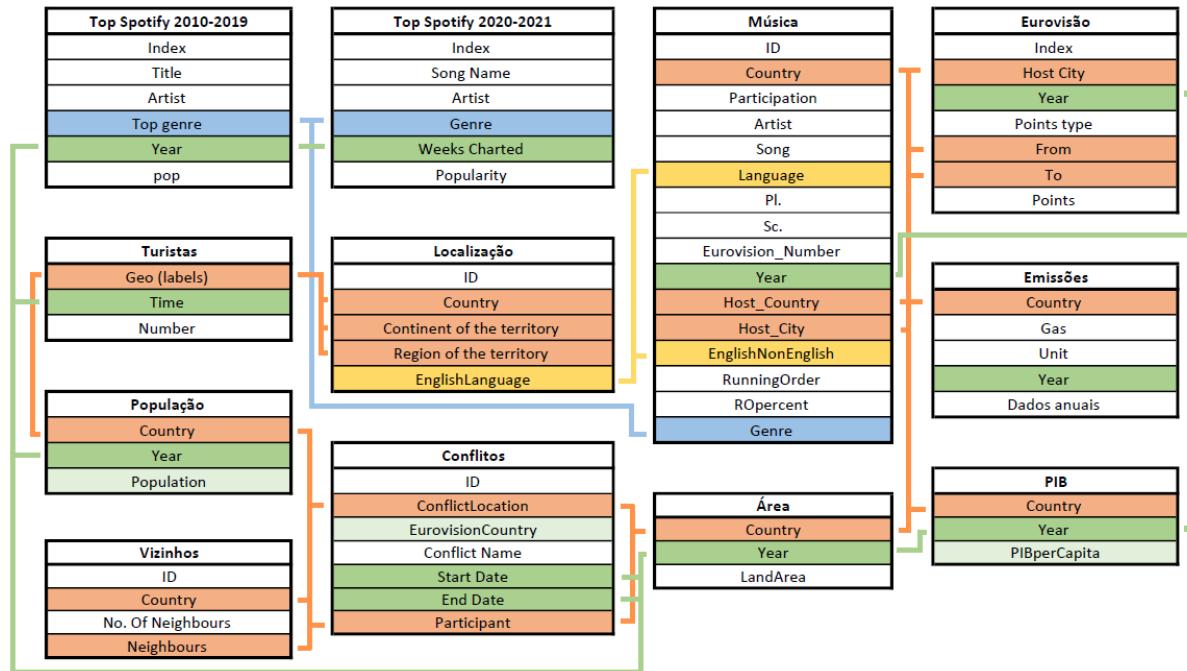


Figura 26 - Diagrama relacional entre tabelas

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 actualidade. 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 está 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		winning chance	BET365	UNIBET	888 SPORT	
1	Ukraine Kalush Orchestra - Stefania	28%	2.37	2.94	2.3		1	Ukraine Kalush Orchestra - Stefania	33%	6/5	6/5
2	Italy Mahmood & Bianco - Brividi	18%	3.5	3.8	3.95		2	Italy Mahmood & Bianco - Brividi	17%	10/3	13/4
3	Sweden Cornelia Jakobs - Hold Me Closer	6%	12	15	11		3	Sweden Cornelia Jakobs - Hold Me Closer	12%	5/1	19/4
4	Greece artist: Amanda Tenfjord	5%	12	15	13		4	United Kingdom Sam Ryder - Space Man	5%	14/1	17/1
5	Poland Krystian Ochman - River	4%	15	24	18		5	Greece Amanda Tenfjord - Die Together	4%	18/1	20/1
6	United Kingdom	4%	17	20	17		6	Poland Ochman - River	4%	18/1	20/1
7	Norway Subwoofler - Give That Wolf a Ban...	3%	23	48	20		7	Spain Chanel - SloMo	2%	25/1	33/1
8	Netherlands S10 - De Diepte	3%	26	32	25		8	Norway Subwoofler - Give That Wolf a Ban...	2%	33/1	30/1
9	Belgium artist: Jérémie Makiese	2%	26	48	25		9	Netherlands S10 - De Diepte	2%	33/1	30/1
10	Spain Chanel - SloMo	2%	34	44	30		10	Australia Sheldon Riley - Not the Same	2%	40/1	50/1
11	Australia Sheldon Riley - Not the Same	2%	29	48	30		11	Portugal Maro - Saudade, saudade	1%	50/1	50/1
12	France Alvan & Ahez - Fulenn	2%	36	44	35		12	Belgium Jérémie Makiese - Miss You	1%	66/1	50/1
13	Switzerland Marius Bear - Boys Do Cry	2%	21	160	30		13	France Alvan & Ahez - Fulenn	1%	66/1	66/1
14	Finland The Rasmus - Jezebel	2%	51	100	40		14	Switzerland Marius Bear - Boys Do Cry	1%	66/1	66/1
15	Cyprus Andromache - Eia	2%	41	65	50		15	Serbia Konstrakta - In Corpore Sano	1%	66/1	60/1

	winning chance	888 SPORT	BET365	UNIBET	LAD BROKES	SMARKETS*	COOL BET		winning chance	BET365	UNIBET	888 SPORT	WILLIAM HILL	COOL BET	BETSSON		
1	Ukraine Kalush Orchestra - Stefania	42%	1.6	1.8	1.6	1.73	1.92	1.93	1	Ukraine Kalush Orchestra - Stefania	60%	1.33	1.34	1.29	1.33	1.4	1.3
2	Italy Mahmood & Bianco - Brividi	15%	5.5	5	5	5	6.6	6	2	Sweden Cornelia Jakobs - Hold Me Closer	11%	6.5	6.75	7.4	7	8	8
3	Sweden Cornelia Jakobs - Hold Me Closer	11%	6.7	7	6.5	7	8.8	6.5	3	United Kingdom Sam Ryder - Space Man	10%	8	9	8.75	7	9	9
4	United Kingdom Sam Ryder - Space Man	7%	9	11	13	11	14	11	4	Italy Mahmood & Bianco - Brividi	5%	15	17	14	12	16	18
5	Spain Chanel - SloMo	4%	17	17	19	19	24	22	5	Spain Chanel - SloMo	5%	15	18	16	15	21	20
6	Greece Amanda Tenfjord - Die Together	3%	22	15	31	26	34	28	6	Serbia Konstrakta - In Corpore Sano	1%	41	51	51	67	91	45
7	Poland Ochman - River	2%	41	41	26	34	75	51	7	Poland Ochman - River	1%	51	71	58	67	91	60
8	Norway Subwoofler - Give That Wolf a Ba...	2%	43	41	31	41	80	51	8	Greece Amanda Tenfjord - Die Together	1%	67	71	69	67	101	70
9	Netherlands S10 - De Diepte	1%	50	51	51	41	120	71	9	Norway Subwoofler - Give That Wolf a Ba...	1%	81	81	73	51	121	75
10	France Alvan & Ahez - Fulenn	1%	77	67	67	41	100	71	10	Netherlands S10 - De Diepte	1%	67	67	76	81	131	75
11	Australia Sheldon Riley - Not the Same	1%	70	71	81	67	130	81	11	Moldova Ziga Zidu & Brothers - Tren...	1%	151	151	201	81	131	150
12	Portugal Maro - Saudade, saudade	1%	73	71	81	67	120	71	12	Finland The Rasmus - Jezebel	<1%	251	251	251	151	201	250
13	Serbia Konstrakta - In Corpore Sano	1%	95	101	81	81	140	91	13	Australia Sheldon Riley - Not the Same	<1%	201	226	251	151	151	200
14	Finland The Rasmus - Jezebel	1%	140	126	81	101	160	101	14	Czech Republic We Are Domi - Lighta...	<1%	251	301	301	151	151	250
15	Switzerland Marius Bear - Boys Do Cry	1%	106	126	81	126	230	201	15	Estonia Stefan - Hope	<1%	251	226	301	251	171	250

Figura 27 - Probabilidade de vencer a Eurovisão em 9/03/2022, 31/03/2022, 01/05/2022 e 14/05/2022 (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? O número de vizinhos de um país tem influência na quantidade de pontos que este 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 público 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]
<b>IDMusica</b>	Chave substituta	ID único gerado sequencialmente	C	1-646	-
<b>NomeMusica</b>	Música cantada na eurovisão	Tabela Eurovision song lyrics, campo Value.Song	T	Ex: Sangen om dig	-
<b>Língua</b>	Língua da música	Tabela Eurovision song lyrics, campo EnglishNonEnglish	C	English, Not_English, Mixed	-
<b>Classificação</b>	Classificação final da música	Tabela Eurovision song lyrics, campo ValuePl	N	1-26	NA (not applicable)
<b>OrdemAtuacao</b>	Número da ordem do concurso em que a música tocou	Tabela Eurovision song lyrics, campo RunningOrder	N	1-27	NQ (not qualified)
<b>Percentagem OrdemAtuacao</b>	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)
<b>Pontuação</b>	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]
IDConflito	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. "Eurovisio nParticipa nt"	
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]
<b>IDEurovisao</b>	Chave Substituta	ID gerado sequencialmente	C	0-64	-
<b>Numero Edicao</b>	Número da edição da Eurovisão	Tabela Música (campo Value.Eurovision_Number)	T	1-65	-
<b>Total Participante</b>	Total de países participantes na edição	Dado gerado com recurso a linguagem de programação	N	0-43	-
<b>Verificacao VotoJuri</b>	Indica se na edição da Eurovisão houve votos do júri	Dado gerado com recurso a linguagem de programação	C	“Jury Vote” “No Jury Vote”	-
<b>Verificacao Televoto</b>	Indica se na edição da Eurovisao houve votos por televoto	Dado gerado com recurso a linguagem de programação	C	“Televote” “No Televote”	-
<b>Votos PorPais</b>	Indica o número total de pontos que um país atribui	Dado gerado com recurso a linguagem de programação	N	6-116	NULL
<b>PontosMax PorPais</b>	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
<b>Descricao Pontos</b>	Campo descritivo dos pontos que cada país pode dar	Dado obtido com informacao da wikipedia (*)	T	96-4988	NULL
<b>Max Pontuacao PorPais</b>	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

<b>Ano</b>	Ano da edição da eurovisão	Tabela Música (campo Value.Eurovision_Number)	D	1956-2021 (1) não há dados de 2020- não se realizou	-
<b>PaisAnfitriao</b>	Pais onde foi realizada a edição da eurovisão	Tabela Música (campo Value.Host_Country)	T	Ex. “Switzerland”	-
<b>CidadeAnfitriap</b>	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
<b>Pontos</b>	Número de pontos que cada música recebeu	Tabela Musica, campo Sc.	0 - 12
<b>PIB per capita</b>	Valor correspondente ao PIB de cada país por ano	Tabela PIB, campo PIBperCapita	1878.1- 36920.8
<b>CO<sub>2</sub></b>	Número de gases anuais emitidos por cada país.	Tabela Emissões, campo Dados Anuais	0.140 - 1790.34
<b>Turistas</b>	Número total de chegadas de turistas a estabelecimento de alojamento turístico por ano por país.	Tabela Turistas, campo Number	56 666 – 140030631
<b>Área</b>	Área total de cada país	Tabela Area campo LandArea	2.027 – 16389950
<b>População</b>	Número de população residente no país	Tabela População campo Population	13410 – 148538197
<b>Densidade populacional</b>	Densidade populacional referente a cada país	Obtida pela divisão da medida de população pela medida de área	
<b>Emissões por população</b>	Emissões de CO <sub>2</sub> por pessoa em cada país	Obtida pela divisão da medida de emissões pela medida de população	
<b>Emissões por área</b>	Emissões de CO <sub>2</sub> por área	Obtida pela divisão da medida de emissões pela medida de área	
<b>Turismo por área</b>	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<sub>2</sub>;
- 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<sup>2</sup>)
- População
- Densidade Populacional
- Emissões por população
- Emissões por população
- Turismo por área

## 6.4 Tabela Multivalor

Após ter sido criada a Dimensão Conflito, deparamo-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.

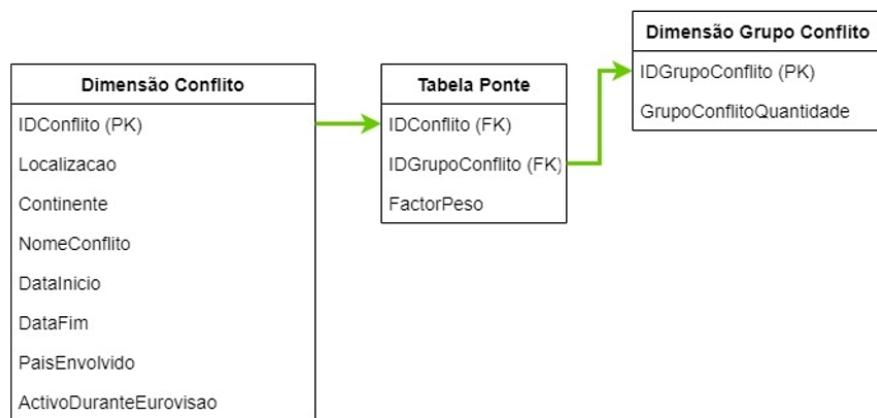


Figura 28 - Tabela Multivalor

## 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.

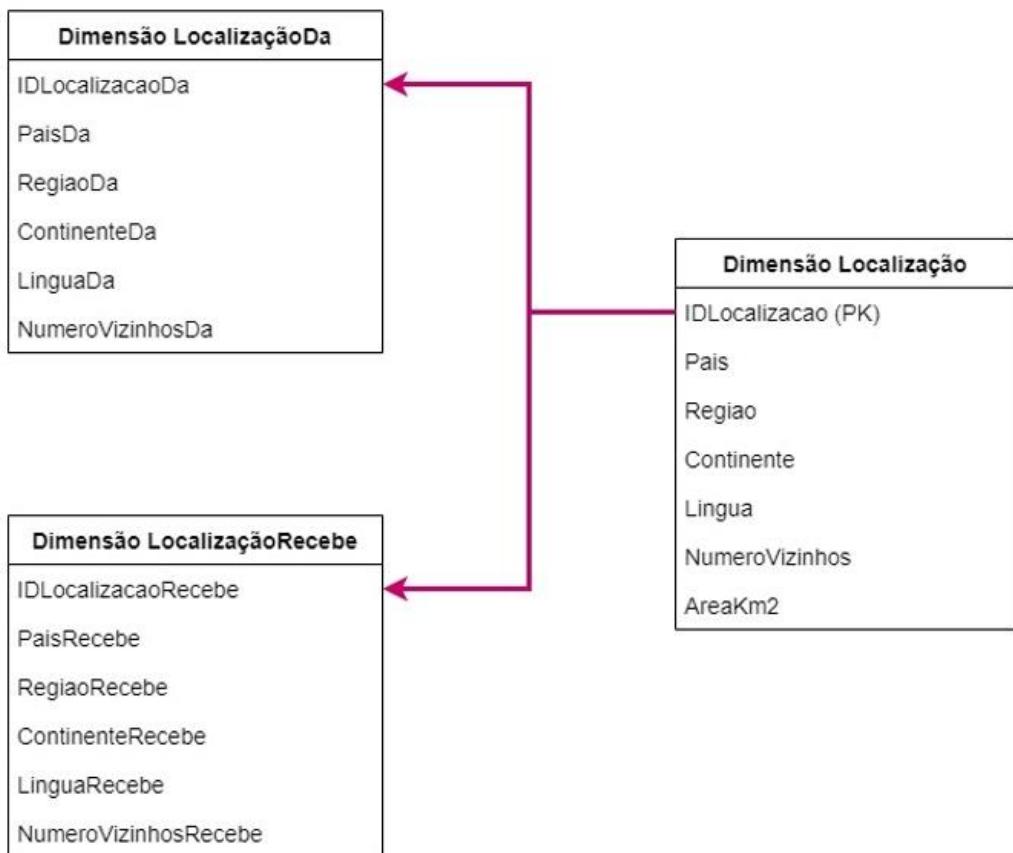


Figura 29: Técnica do role-playing

## 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.

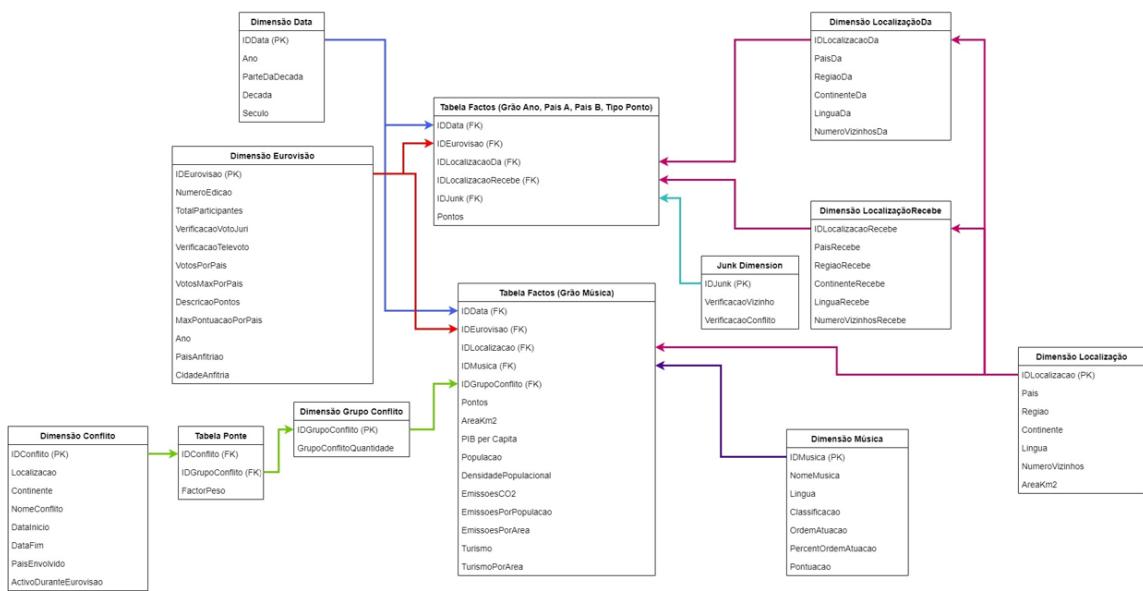


Figura 30- Esquema em estrela global

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

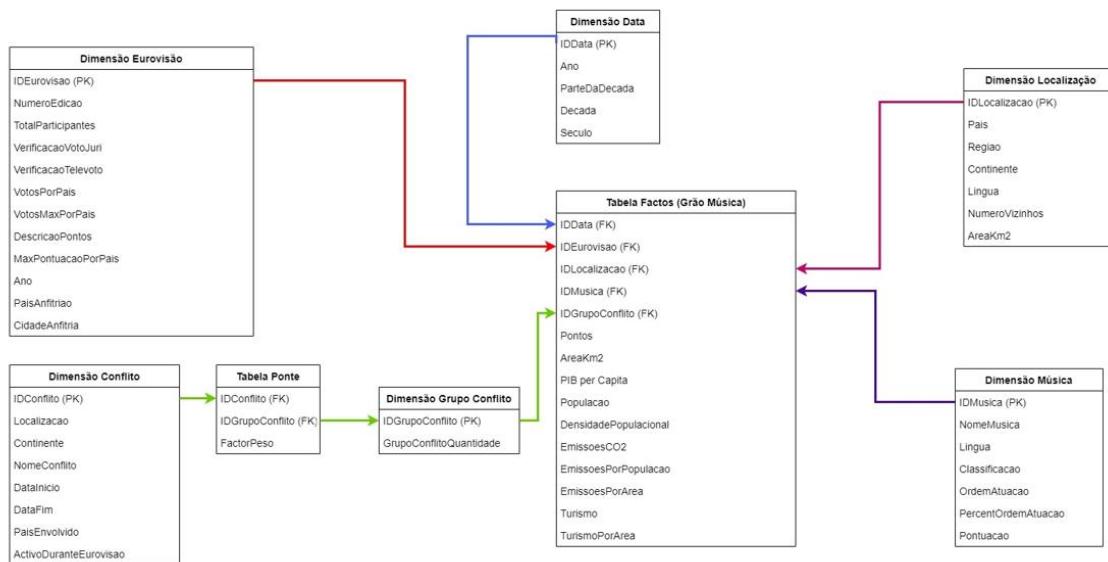


Figura 31 - Tabela de Factos grão Música

Na Figura 32 observa-se as ligações entre as dimensões e a tabela de factos com um grão por país, país, tipo, ano.

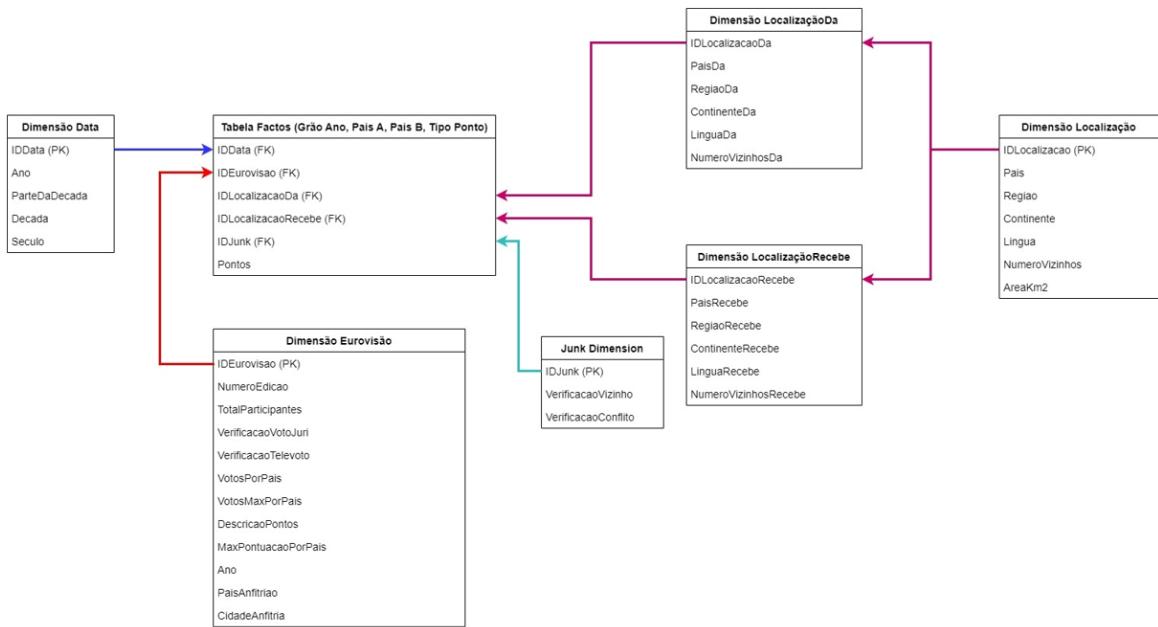


Figura 32 - Tabela de factos grão País, País, Tipo, Ano

## 7. Desenvolvimento dos Programas que Compõe o Sistema ETL

Para a fase inicial da criação do sistema ETL optou-se pela utilização de ficheiros de texto (formato .csv) na *data staging area* em vez da construção de uma base de dados devido à maior facilidade de tratar dados com recurso a linguagem *Python* e à biblioteca *Pandas*. O processo de criação das dimensões e das tabelas de factos é descrito em seguida.

### 7.1 Criação das dimensões com recurso a Python

As dimensões Data, JunkDimension e ConflitosSimplificado são dimensões que não têm dataset de origem e por isso foram geradas de raiz para esta fase. Para a dimensão Data foi identificado o intervalo de tempo que corresponde aos nossos dados durante a primeira fase do projeto e foram criadas outras medidas relevantes a partir do campo “ano” que podem ser utilizadas como hierarquia ou para obter resultados com maior granularidade.

```

#Construcao da Dimensao Data
#Dimensao Data nao tem tabela de origem - feita de raiz no python com intervalo de tempo correspondente ao identificado na fase1
#Novo Dataframe
dfdata=pd.DataFrame(columns=['IDData','Ano'])

#Criacao da Coluna Ano com 66 periodos e frequencia anual
dfdata['Ano']=pd.period_range("1/1/1956", freq="Y", periods=66)

#Criacao da coluna ID atraves do uso do indice do pandas
dfdata['IDData']=dfdata.index*1

#Passar Ano do tipo range para tipo int de forma a poder fazer operacoes sobre a coluna
dfdata['Ano'] = dfdata['Ano'].astype(str).astype(int)

#Criacao da coluna Metade da Decada atraves de condicoes com recurso ao resto inteiro da divisao por 10
conditions2=[((dfdata['Ano']%10)<5),(dfdata['Ano']%10)>=5]
values2=['Primeira Metade da Decada','Segunda Metade da Decada']
dfdata['MetadeDaDecada']=np.select(conditions2,values2)

#Criacao da coluna Decada com condicoes de intervalo
conditions=[(dfdata['Ano']<1960),
            (dfdata['Ano']>=1960) & (dfdata['Ano']<1970),
            (dfdata['Ano']>=1970) & (dfdata['Ano']<1980),
            (dfdata['Ano']>=1980) & (dfdata['Ano']<1990),
            (dfdata['Ano']>=1990) & (dfdata['Ano']<2000),
            (dfdata['Ano']>=2000) & (dfdata['Ano']<2010),
            (dfdata['Ano']>=2010) & (dfdata['Ano']<2020),
            (dfdata['Ano']>=2020)]
values=[1950,1960,1970,1980,1990,2000,2010,2020]
dfdata['Decada']=np.select(conditions,values)

#Criacao da coluna Seculo com condicoes de intervalo
conditions1=[(dfdata['Ano']<2000),dfdata['Ano']>=2000]
values1=[20,21]
dfdata['Seculo']=np.select(conditions1,values1)

#Resultado Final
dfdata.to_csv("DimData_final.csv", index=False)

```

Figura 33 - Código Python para Criação da Dimensão Data

As dimensões de JunkDimension e ConflitosSimplificado (também chamada JunkConflitos) são dimensões que servem como validação de outros campos e foram, também, criadas de raiz, sendo que a tabela JunkDimension valida dados relativos a países vizinhos e tipos de voto e a tabela ConflitosSimplificado valida dados relativamente à existência de conflitos, à sua localização e à sua existência temporal durante o festival da Eurovisão. Embora esta segunda tenha sido gerada de raiz no programa, ela necessita da Dimensão Conflito, que não vai entrar no modelo final, para fazer validação dos dados com a Tabela de Factos.

```

#Construcao da Junk Dimension - sem dataset de origem

#Novo dataframe
dfjunk=pd.DataFrame(columns=['IDJunk','VerificacaoVizinho','VerificacaoVoto'])

#Criacao de valores de colunas
dfjunk['VerificacaoVizinho']=['PaisesVizinhos','PaisesVizinhos','PaisesVizinhos','PaisesNaoVizinhos','PaisesNaoVizinhos']
dfjunk['VerificacaoVoto']=['Televoto','VotoJuri','VotoNaoDiscriminado','Televoto','VotoJuri','VotoNaoDiscriminado']

#Criacao da coluna ID atraves do uso do indice do pandas
dfjunk['IDJunk']=dfjunk.index+1

#Resultado Final
dfjunk.to_csv("DimJunk_final.csv", index=False)

#Construcao da Dimensao Conflitos Simplificado
#Esta dimensao funciona como uma junk dimension

#Novo dataframe
dfconflito=pd.DataFrame(columns=['IDConflito','VerificacaoConflito','LocalizacaoConflito'])

#Criacao de valores de colunas
dfconflito['VerificacaoConflito']=['NaoEnvolvidoEmConflito','EnvolvidoEmConflito','EnvolvidoEmConflito','EnvolvidoEmConflito','EnvolvidoEmConflito']
dfconflito['LocalizacaoConflito']=['NA','ConflitoNumParticipanteEurovisao','ConflitoNumParticipanteEurovisao','ConflitoNumNaoParticipanteEurovisao','AtivoEurovisao']
dfconflito['AtivoEurovisao']=['NA','AtivoDuranteEurovisao','NaoAtivoDuranteEurovisao','AtivoDuranteEurovisao','NaoAtivoDuranteEurovisao']

#Criacao da coluna ID atraves do uso do indice do pandas
dfconflito['IDConflito']=dfconflito.index+1

#Resultado Final
dfconflito.to_csv("DimJunkConflito_final.csv", index=False)

```

Figura 34 - Código Python para Criação das Dimensões Junk

As dimensões Música e Localização foram criadas com base em tabelas de datasets obtidos durante a primeira fase do trabalho. Foi também criada a Dimensão Conflitos, embora esta não tenha sido utilizada no trabalho final, visto que esta é necessária para associar as chaves da Dimensão ConflitosSimplificado à tabela de factos.

```

#Construcao da Dimensao Localizacao
#2 Dataset origem: Localizacao e Vizinhos
dflocal_1=pd.read_excel('Localizacao.xlsx')
dfvizinhos=pd.read_excel('Vizinhos.xlsx', sheet_name='LB_Eurovision')

#Novo dataframe
dflocal=pd.DataFrame(columns=['IDLocalizacao','Pais'])

#Alteracao de nome da coluna para facilitar o merge
dfvizinhos.rename(columns = {'Name':'English Name'}, inplace = True)

#Eliminacao de colunas que nao vao ser utilizadas
dfvizinhos.drop(columns=['Neighbour10', 'Neighbour9', 'Neighbour8', 'Neighbour7', 'Neighbour6', 'Neighbour5', 'Neighbour4', 'Neighbour3'])

#Merge dos dois dataset de origem
result = pd.merge(dflocal_1, dfvizinhos, on=["English Name"])

#Composicao da dimensao Localizacao
dflocal['Pais']=result['English Name']
dflocal['Region']=result['Region of the territory']
dflocal['Continente']=result['Continent of the territory']
dflocal['IDLocalizacao']=dflocal.index+1
dflocal['NumeroDeVizinhos']=result['NrNeighboursinEurope']

#Resultado Final
dflocal.to_csv("DimLocalizacao_final.csv", index=False)

```

Figura 35 - Código Python para Criação da Dimensão Localização

```

#Construcao da Dimensao Conflitos
#Esta dimensao nao foi utilizada diretamente no data warehouse mas utilizada para verificar os ID da JunkConflict utilizados

#Esta dimensao utiliza dois dataset de origem: Localizacao e Conflitos. Como a dimensao Localizacao tinha sido ja criada
#optamos por utilizar esta em vez do dataset original pois continha menos dados que nao eram uteis facilitando a limpeza
dfl=pd.read_csv('DimLocalizacao.csv')
dfc=pd.read_excel('Conflitos.xlsx')
dfconflito=pd.DataFrame(columns=['IDConflito'])

#Criação das colunas de importação direta dos datasets
dfconflito['LocalizacaoConflito']=dfc['Conflict Location']
dfconflito['IDConflito']=dfconflito.index+1
dfconflito['ParticipaEurovisao']=np.where(dfconflito['LocalizacaoConflito'].isin(dfl['País']), 'ParticipanteEurovisao', 'NãoParticipante')
dfconflito['NomeConflito']=dfc['Conflict']
dfconflito['DataInicio']=dfc['Start date']
dfconflito['DataFim']=dfc['End date']

#Passagem das datas para Ano de forma a respeitar a granularidade da TF
dfconflito['AnoInicio']=pd.DatetimeIndex(dfconflito['DataInicio']).year
dfconflito['AnoFim']=pd.DatetimeIndex(dfconflito['DataFim']).year

#Continuação da criação das colunas
dfconflito['Participante']=dfc['Participant 1']

#Quando o conflito ainda não está concluído, toma o valor de 9999 para ser possível passar o tipo da coluna para inteiro
#e realizar operações numéricas
dfconflito['AnoFim'].fillna('9999', inplace=True)
dfconflito.AnoFim = dfconflito.AnoFim.astype(int)

#Criação de condições para o conflito ativo ou não durante a eurovisão
conditions=[(dfconflito['DataFim']-dfconflito['DataInicio']).dt.days>365,
            (pd.DatetimeIndex(dfconflito['DataInicio']).year==(pd.DatetimeIndex(dfconflito['DataFim']).year)) & (pd.DatetimeIndex(dfconflito['DataInicio']).year==(pd.DatetimeIndex(dfconflito['DataFim']).year)) & (pd.DatetimeIndex(dfconflito['DataInicio']).year==(pd.DatetimeIndex(dfconflito['DataFim']).year)) & (pd.DatetimeIndex(dfconflito['DataInicio']).year==(pd.DatetimeIndex(dfconflito['DataFim']).year)) & (pd.DatetimeIndex(dfconflito['DataInicio']).year!=pd.DatetimeIndex(dfconflito['DataFim']).year) & (pd.DatetimeIndex(dfconflito['DataInicio']).year!=pd.DatetimeIndex(dfconflito['DataFim']).year) & (pd.DatetimeIndex(dfconflito['DataInicio']).year!=pd.DatetimeIndex(dfconflito['DataFim']).year) & (pd.DatetimeIndex(dfconflito['DataFim']).year==9999)]
values=['AtivoDuranteEurovisao', 'AtivoDuranteEurovisao', 'NãoAtivoDuranteEurovisao', 'NãoAtivoDuranteEurovisao', 'AtivoDuranteEurovisao', 'EstadoDuranteEurovisao']=np.select(conditions, values)

#Alterar os valores de 9999 para vazio de forma a facilitar a importação no Postgresql
dfconflito['AnoFim'] = dfconflito.AnoFim.replace(9999, '')

#Apagar as colunas de data que já não são necessárias
dfconflito=dfconflito.drop(columns=['DataInicio', 'DataFim'])

#Resultado Final
dfconflito.to_csv('DimConflito_final.csv')

```

Figura 36 - Código Python para Criação da Dimensão Conflitos

```

#Construcao da Dimensao Musica
#Esta dimensao teve como tabela de origem o dataset musica

df = pd.read_excel('Musica.xlsx')
dfmusica=pd.DataFrame(columns=['IDMusica','Musica'])

#Preenchimento dos campos de importação direta
dfmusica['Musica']=df['Value.Song']
dfmusica['IDMusica']=dfmusica.index+1
dfmusica['Lingua']=df['EnglishNonEnglish']
dfmusica['Classificacao']=df['Value.Pl.']

#Alterar os valores de NaN para vazio de forma a facilitar a importação no Postgresql
dfmusica.fillna('', inplace=True)

#Preenchimento dos campos de importação direta
dfmusica['OrdemAtuacao']=df['RunningOrder']

#Alterar os valores de NaN para vazio de forma a facilitar a importação no Postgresql
dfmusica.fillna('', inplace=True)

#Preenchimento dos campos de importação direta
dfmusica['OrdemAtuacaoNormalizada']=df['Ropercents']
dfmusica['Pontuacao']=df['Value.Sc.']

#Alterar os valores de NaN para vazio de forma a facilitar a importação no Postgresql
dfmusica.fillna('', inplace=True)

#Resultado final
dfmusica.to_csv("DimMusica_final.csv", index=False)

```

Figura 37 - Código Python para Criação da Dimensão Música

A última dimensão que é necessária para o projeto é a dimensão Eurovisão que, como referido anteriormente, tinha sido criada por nós originalmente no Excel. Infelizmente não foi possível recriar esta tabela no *Python* devido ao facto de ela conter informação que tinha sido pesquisada posteriormente e que não estava incluída nos datasets originais. Assim sendo, optámos por manter a tabela que tinha sido criada no Excel e não tentar a sua criação através de código de forma a otimizar a nossa gestão de tempo.

## 7.2 Manipulação dos datasets de medidas com recurso a *Python*

Após serem criadas todas as dimensões, foi necessário tratar os dados que estavam nos datasets com as medidas consideradas para a segunda tabela de factos. Uma vez que os datasets Turistas, Emissões, PIB, População e Área estavam dispostos no formato horizontal, exemplificado na figura 38, recorreu-se à linguagem *Python* para realizar a transformação destes dados num formato mais acessível para a criação da tabela de factos. Um exemplo do código utilizado para este processo encontra-se na figura 39 e o resultado é visível na figura 40.

	Country	2018	2017	2016	2015	2014	2013	2012	2011	2010	...	1999	1998	1997	1996	1995	1994
0	Albania	5.32	5.37	4.70	4.23	4.37	3.95	3.74	4.20	3.46	...	3.16	1.99	1.67	2.16	2.14	2.33
1	Andorra	0.46	0.47	0.47	0.47	0.46	0.48	0.49	0.49	0.52	...	0.51	0.49	0.47	0.45	0.43	0.41
2	Armenia	5.59	5.36	5.10	5.26	5.40	5.42	5.64	4.86	4.46	...	3.24	3.60	3.47	2.71	3.64	2.87
3	Australia	388.81	389.39	386.15	346.53	340.43	349.83	356.33	356.33	416.72	...	383.46	378.07	357.18	349.99	339.40	329.03
4	Austria	56.06	58.71	56.27	56.08	54.92	58.68	58.04	61.20	63.79	...	49.58	51.20	50.75	51.43	47.71	44.81
5	Belarus	37.80	34.80	33.85	38.47	43.05	43.53	43.38	41.94	36.42	...	22.42	24.49	26.24	25.62	24.64	31.66
6	Belgium	93.63	92.93	94.48	92.84	87.80	94.25	92.60	93.21	103.46	...	111.59	115.49	112.72	115.53	110.65	110.99
7	Bosnia and Herzegovina	20.81	20.91	20.60	15.90	15.77	18.13	18.25	20.02	21.19	...	13.77	14.02	11.75	7.44	6.63	6.42

Figura 38 - Formato inicial do dataset

```
In [ ]: import psycopg2 as pg
import pandas as pd
import os

In [ ]: #Read file
df=pd.read_excel("emissoes.xlsx")
df

In [ ]: df1 = pd.melt(df, id_vars = ["Country"])
df1

In [ ]: df1['value'] = df1.value.replace(':', '')

In [ ]: df1.to_csv('Emissoes.csv')
```

Figura 39 - Código Python para Manipulação dos Datasets de Medidas

	Country	variable	value
0	Albania	2018	5.32
1	Andorra	2018	0.46
2	Armenia	2018	5.59
3	Australia	2018	388.81
4	Austria	2018	56.06
...	...	...	...
1416	Azerbaijan	1990	55.94
1417	Georgia	1990	15.22
1418	Israel	1990	33.78
1419	Morocco	1990	16.83
1420	Macedonia	1990	8.59

Figura 4o - Formato resultante do código Python aplicado ao Dataset

### 7.3 Criação da tabela de factos 1

Em primeiro lugar, foi criada a tabela de factos com granularidade (país dá, país recebe, ano, tipo de pontos) pois esta foi baseada quase por inteiro numa tabela que tinha sido extraída na primeira fase do projeto, tendo sido apenas necessário fazer a implementação das chaves substitutas e a verificação dos critérios de identificação para a JunkDimension.

```

#Alteracao do nome da coluna para possibilitar a existencia de duas chaves estrangeiras
dfbase.rename(columns = {'IDLocalizacao':'IDLocalizacaoDa'}, inplace = True)

#Apagar colunas que nao irao ser utilizadas
dfbase=dfbase.drop(['Pais','Regiao','Continente','NumeroDeVizinhos'],axis=1)

#Segundo merge com a dimLocalizacao e alteracao do nome da coluna
dfbase = pd.merge(dfbase, df2, how='outer', left_on = 'To', right_on = 'Pais')
dfbase.rename(columns = {'IDLocalizacao':'IDLocalizacaoRecebe'}, inplace = True)

#Apagar colunas que nao irao ser utilizadas
dfbase=dfbase.drop(['Pais','Regiao','Continente','NumeroDeVizinhos'],axis=1)
|
#Juncao com a tabela de vizinhos
dfbase = pd.merge(dfbase, df5, how='outer', left_on = 'From', right_on = 'Name')

#Procura de paises vizinhos
conditions2=[(dfbase['To']==dfbase['Neighbour1']),
            (dfbase['To']==dfbase['Neighbour2']),
            (dfbase['To']==dfbase['Neighbour3']),
            (dfbase['To']==dfbase['Neighbour4']),
            (dfbase['To']==dfbase['Neighbour5']),
            (dfbase['To']==dfbase['Neighbour6']),
            (dfbase['To']==dfbase['Neighbour7']),
            (dfbase['To']==dfbase['Neighbour8']),
            (dfbase['To']==dfbase['Neighbour9']),
            (dfbase['To']==dfbase['Neighbour10'])]

values2=['PaisesVizinhos','PaisesVizinhos','PaisesVizinhos','PaisesVizinhos','PaisesVizinhos',
        'PaisesVizinhos','PaisesVizinhos','PaisesVizinhos','PaisesVizinhos','PaisesVizinhos']
dfbase['VerifVizinho']=np.select(conditions2,values2)

#definir todos os vazios da coluna como paises nao vizinhos
vizinho_col = dfbase['VerifVizinho']
vizinho_col.replace(to_replace = '0', value = 'PaisesNaoVizinhos', inplace=True)

#Alteracao de valores na tabela Points type para fazer juncao com a Junk Dimension
ptype_col = dfbase['Points type']
ptype_col.replace(to_replace = 'Points given by televoters', value = 'Televoto', inplace=True)
ptype_col.replace(to_replace = 'Points given by the jury', value = 'VotoJuri', inplace=True)
ptype_col.replace(to_replace = 'Points given', value = 'VotoNaoDiscriminado', inplace=True)

#Mudanca do tipo de dados das colunas para permitir o merge dos dois dataframes
dfbase['Points type'] = dfbase['Points type'].astype(str)
dfbase['VerifVizinho'] = dfbase['VerifVizinho'].astype(str)
dfjunk['VerificacaoVizinho']=dfjunk['VerificacaoVizinho'].astype(str)
dfjunk['VerificacaoVoto']=dfjunk['VerificacaoVoto'].astype(str)

#Merge com a JunkDimension
dfbase = pd.merge(dfbase, dfjunk, how='left', left_on = ['Points type','VerifVizinho'],
                  right_on = ['VerificacaoVoto','VerificacaoVizinho'])

```

Figura 41 – Exerto do Código Python para Criação da Tabela de Factos 1

As principais dificuldades na criação desta tabela foram o aparecimento de linhas vazias no dataframe e a manutenção dos domínios dos campos do dataframe, visto que nesta biblioteca não são aceites valores nulos “NaN” como tipo inteiro, pelo que tiveram de ser feitas várias validações sobre os dados de forma a garantir que os domínios de cada coluna estavam corretos.

#### 7.4 Criação da tabela de factos 2

A segunda tabela de factos, com granularidade música, foi criada de seguida. Esta necessitou de um maior número de juncões com outras tabelas, assim como verificações para ligar com a tabela ConflitosSimplificado.

Nesta tabela foram também pré-calculados alguns resultados sobre as medidas numéricas de forma a facilitar a sua utilização nos relatórios.

```

#Apagar as colunas do dataset base que nao sao relevantes para a tabela de factos
dfbase=dfbase.drop(['Value.#','Value.Artist','Genre','Value.Language','EnglishNonEnglish',
                    'Value.PL','RunningOrder','Value.Host_Country','Value.Host_City',
                    'Value.Lyrics','Value.Lyrics translation','Ropercent','Value.#.1'],axis=1)
dfbase.rename(columns = {'Value.Sc.':'Pontuacao'}, inplace = True)

#Merge com os datasets que contem as medidas numericas na seguinte ordem:
#1-Merge, 2-Renomear coluna que vai ser mantida, 3-Apagar colunas nao necessarias
#Dataset Populacao
dfbase = pd.merge(dfbase, dfpop, how='left', left_on = ['Value.Country','Value.Year'] , right_on = ['Country','variable'])
dfbase.rename(columns = {'value':'Populacao'}, inplace = True)
dfbase=dfbase.drop(['Unnamed: 0','Country','variable'],axis=1)

#Dataset Area
dfbase = pd.merge(dfbase, dfarea, how='left', left_on = ['Value.Country','Value.Year'] , right_on = ['Country Name','variable'])
dfbase.rename(columns = {'value':'AreaKM2'}, inplace = True)
dfbase=dfbase.drop(['Unnamed: 0','Country Name','variable'],axis=1)

#Dataset PIB
dfbase = pd.merge(dfbase, dfgdp, how='left', left_on = ['Value.Country','Value.Year'] , right_on = ['Country Name','Year'])
dfbase=dfbase.drop(['Unnamed: 0','Country Name','Year','id'],axis=1)

#Dataset Turistas
dfbase = pd.merge(dfbase, dfturist, how='left', left_on = ['Value.Country','Value.Year'] , right_on = ['TIME','variable'])
dfbase=dfbase.drop(['Unnamed: 0','TIME','variable'],axis=1)
dfbase.rename(columns = {'value':'Turistas'}, inplace = True)

#Dataset Emissoes
dfbase = pd.merge(dfbase, dfco2, how='left', left_on = ['Value.Country','Value.Year'] , right_on = ['Country','variable'])
dfbase=dfbase.drop(['Unnamed: 0','Country','variable'],axis=1)
dfbase.rename(columns = {'value':'EmissoesCO2'}, inplace = True)

#Merge com os datasets que vao necessitar de criacao de chaves substitutas e eliminacao de colunas desnecessarias
#Merge com dimData
dfbase = pd.merge(dfbase, dfdata, how='outer', left_on = 'Value.Year', right_on = 'Ano')
dfbase=dfbase.drop(['MetadeDaDecada','Decada','Seculo','Ano'],axis=1)
#Merge com dimLocalizacao
dfbase = pd.merge(dfbase, dflocal, how='outer', left_on = 'Value.Country', right_on = 'Pais')
dfbase=dfbase.drop(['NumeroDeVizinhos','Continente','Regiao','Pais'],axis=1)
#Merge com dimEurovisao
dfbase = pd.merge(dfbase, dfeurovisao, how='outer', left_on = 'Value.Year', right_on = 'Ano')
dfbase=dfbase.drop(['Regralinguagem','CidadeAnfitria','PaisAnfitriao','Ano','MaxPontuacaoPais','TotalPontosDisponivelConcurso',
                    'TotalPontosPaisDa','MaxPontosPaisADaPaisB','DescricaoPontos','VerificacaoTelevoto','VerificacaoVotoJuri',
                    'TotalParticipantes','NumeroEdicao'],axis=1)
#Merge com dimMusica
dfbase = pd.merge(dfbase, dfmusica, how='outer', left_on = 'Value.Song', right_on = 'Musica')
dfbase=dfbase.drop(['Pontuacao_Y','OrdemAtuacaoNormalizada','OrdemAtuacao','Classificacao','Lingua','Musica'],axis=1)

```

Figura 42 - Excerto do Código Python para Criação da Tabela de Factos 2

No caso desta segunda tabela de factos, o principal desafio foi a quantidade enorme de colunas que se iam adicionando à medida que se faziam juncções entre a tabela de facto e as dimensões, pelo que se adotou uma estratégia de eliminar as colunas que não eram relevantes imediatamente a seguir à juncão. Tal como na outra tabela de factos, existiu alguma dificuldade na manutenção das dimensões das chaves devido aos valores ‘NaN’, em particular no caso das músicas que pertenciam a 2020, e que, de forma a evitar apagar dados, foram colocadas numa Eurovisão fictícia com código 99 e descrição de festival cancelado.

## 7.5 Implementação das tabelas de factos no PostgreSQL e PowerBI

Depois de criadas as dimensões e as tabelas de factos através do código *Python*, procedeu-se à sua implementação no programa *PostgreSQL*.

Para tal, utilizaram-se comandos SQL para criação das tabelas e definição de chaves primárias e estrangeiras, de modo a garantir a ligação e integridade dos dados, e algumas verificações no formato de “checks”, em especial para campos com valores categóricos, para assegurar que estes correspondiam aos valores esperados e não permitir a existência de erros nos dados que vão depois ser utilizados nos relatórios. As tabelas foram depois importadas para o *PostgreSQL* com a opção *import* incluída no programa.

```

CREATE TABLE PUBLIC.DimData(
IDData      INT PRIMARY KEY,
Ano         INT,
ParteDaDecada  VARCHAR(35) CHECK (ParteDaDecada IN ('Segunda Metade da Decada', 'Primeira Metade da Decada')),
Decada      INT,
Seculo       INT CHECK (Seculo IN (20,21));

CREATE TABLE PUBLIC.DimLocalizacao(
IDLocalizacao  INT PRIMARY KEY,
País          VARCHAR(50),
Região        VARCHAR(50),
Continente    VARCHAR(15) CHECK (Continente IN ('Europa','Asia','Oceania','Africa')),
NúmeroVizinhos INT
);

CREATE TABLE PUBLIC.Musica(
IDMusica      INT PRIMARY KEY,
Música        VARCHAR(300),
Língua        VARCHAR(100) CHECK (Língua IN ('English','NotEnglish','Mixed')),
Classificação INT,
OrdemAtuação   INT,
OrdemAtuaçãoNormalizada FLOAT     CHECK (OrdemAtuaçãoNormalizada<=1),
Pontuação     INT
);

```

Figura 43 – Excerto dos comandos para criação das dimensões no PostgreSQL

```

CREATE TABLE PUBLIC.FactTable1 (
IDData      INT,
IDEurovisão  INT,
IDLocalização  INT,
IDLocalizaçãoRecebe INT,
IDJunk       INT,
Pontos      INT,
CONSTRAINT pk_tfatos1 PRIMARY KEY (IDData, IDLocalizaçãoDa, IDLocalizaçãoRecebe, IDJunk)),
CONSTRAINT fk_tfatos1_data FOREIGN KEY (IDData) REFERENCES PUBLIC.DimData(IDData),
CONSTRAINT fk_tfatos1_euro FOREIGN KEY (IDEurovisão) REFERENCES PUBLIC.DimEurovisão(IDEurovisão),
CONSTRAINT fk_tfatos1_local FOREIGN KEY (IDLocalizaçãoDa) REFERENCES PUBLIC.DimLocalização(IDLocalização),
CONSTRAINT fk_tfatos1_loca2 FOREIGN KEY (IDLocalizaçãoRecebe) REFERENCES PUBLIC.DimLocalização(IDLocalização),
CONSTRAINT fk_tfatos1_junk FOREIGN KEY (IDJunk) REFERENCES PUBLIC.DimJunk(IDJunk)
);

CREATE TABLE PUBLIC.FactTable2 (
IDData      INT NOT NULL,
IDEurovisão  INT NOT NULL,
IDLocalização  INT NOT NULL,
IDMúsica     INT PRIMARY KEY,
IDConflito    INT NOT NULL,
Pontos      INT NOT NULL,
ÁreaKM2      FLOAT,
PIBPerCapita  FLOAT,
População    FLOAT,
DensidadePopulacional  FLOAT,
EmissõesCO2   FLOAT,
EmissõesPorPopulação  FLOAT,
EmissõesPorÁrea  FLOAT,
Turismo       FLOAT,
TurismoPorÁrea  FLOAT,
CONSTRAINT fk_tfatos2_data FOREIGN KEY (IDData) REFERENCES PUBLIC.DimData(IDData),
CONSTRAINT fk_tfatos2_euro FOREIGN KEY (IDEurovisão) REFERENCES PUBLIC.DimEurovisão(IDEurovisão),
CONSTRAINT fk_tfatos2_loca FOREIGN KEY (IDLocalização) REFERENCES PUBLIC.DimLocalização(IDLocalização),
CONSTRAINT fk_tfatos2_musi FOREIGN KEY (IDMúsica) REFERENCES PUBLIC.DimMúsica(IDMúsica),
CONSTRAINT fk_tfatos2_conf FOREIGN KEY (IDConflito) REFERENCES PUBLIC.DimJunkConflito(IDConflito)
);

```

Figura 44 - Criação das tabelas de factos no PostgreSQL

Finalmente, foram também criadas as vistas materializadas sobre a dimensão localização (*roleplaying*) para facilitar a visualização desta informação.

```

CREATE MATERIALIZED VIEW DimLocalizacaoDa
AS
    SELECT l.idlocalizacao AS IDLocalizacaoDa, l.pais AS PaisDa, l.regiao AS RegiaoDa,
    l.continente AS ContinenteDa, l.numerovizinhos AS NumeroVizinhosDa
    FROM dimlocalizacao l
WITH NO DATA;

REFRESH MATERIALIZED VIEW DimLocalizacaoDa;

CREATE MATERIALIZED VIEW DimLocalizacaoRecebe
AS
    SELECT l.idlocalizacao AS IDLocalizacaoRecebe, l.pais AS PaisRecebe, l.regiao AS RegiaoRecebe,
    l.continente AS ContinenteRecebe, l.numerovizinhos AS NumeroVizinhosRecebe
    FROM dimlocalizacao l
WITH NO DATA;

REFRESH MATERIALIZED VIEW DimLocalizacaoRecebe;

```

Figura 45 - Criação das vistas materializadas no PostgreSQL

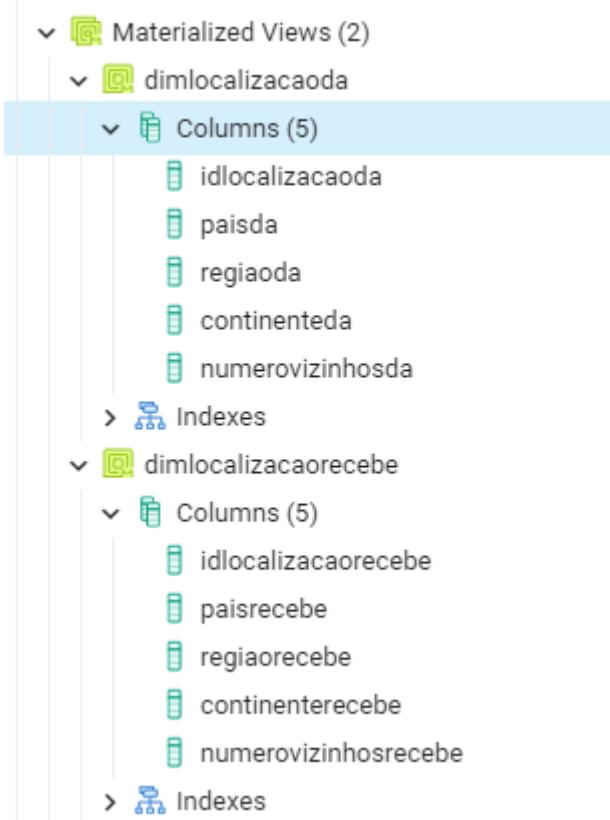


Figura 46 - Vistas implementadas no PostgreSQL

Nesta fase foram detetados alguns erros que foram posteriormente corrigidos no código *Python* ou nos datasets originais para gerar novos ficheiros. Um desses erros foi, por exemplo, a utilização apenas do título da música na dimensão Música, que não é suficiente para gerar uma chave única pois existem 81 instâncias de músicas com títulos repetidos. Assim foi necessário incluir o artista para este par ser único e permitir o campo ID único. Por outro lado, foi também descoberto que os dados referentes à Áustria e Austrália estavam misturados, o que resultava em chaves repetidas. Neste caso, os dados do dataset original tiveram de ser validados com recurso ao site oficial da Eurovisão.

Com o intuito de conseguir ter uma melhor visualização para os resultados da prospeção de dados que visam dar resposta às nossas questões analíticas, propostas na primeira fase do projeto, as tabelas foram também importadas para o programa *PowerBI*.

## 8. Responsabilidades, Inputs e Outputs dos Programas

O seguinte esquema demonstra os ficheiros de entrada, ficheiros de saída e as operações realizadas em *python* para cada um dos datasets no processo de transformá-los nas dimensões, assim como os datasets e dimensões utilizados para a criação de cada uma das tabelas de factos.

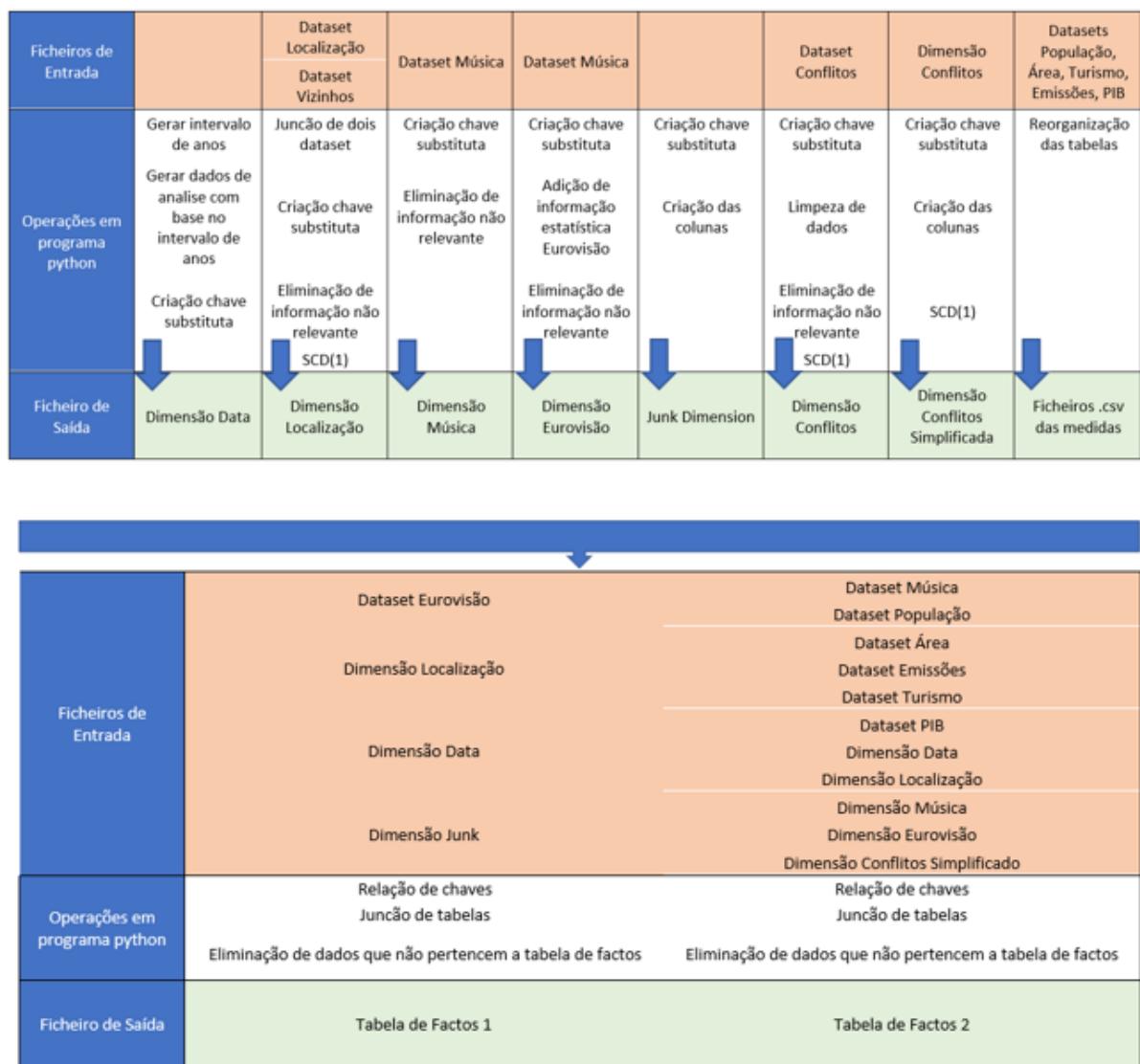


Figura 47 - Diagrama dos programas do sistema ETL

## 9. Desenho do Diagrama do Sistema ETL

O sistema ETL é composto por três fases: em primeiro lugar uma fase de extração, seguido de uma fase de transformação e finalmente a fase de carregamento.

Numa primeira etapa os dados foram extraídos depois de pesquisa em diversos sites que fornecem dados abertos, como o *Kaggle* ou o portal de dados abertos da Comissão Europeia. Nesta fase foi decidido o tema do projeto bem como a aplicação prática a nível das casas de apostas, sendo que esta aplicação nos levou a procurar dados que permitissem procurar uma relação entre os dados históricos do festival da Eurovisão, através da caracterização de alguns aspectos dos países que nele participam. O foco do trabalho foi então encontrar dados que se distribuíssem numa dimensão espacial e temporal, assim como características intrínsecas dos países que são imutáveis ou de mudança lenta. No final da fase de extração foram obtidos 12 datasets.

Na segunda etapa do sistema os dados são trabalhados e transformados na *data staging area*. O objetivo desta etapa é a normalização e conformação dos dados, o que inclui validação da informação e limpeza de erros. Inicialmente os ficheiros foram trabalhados apenas com Excel, sendo as colunas validadas manualmente e novos campos gerados apenas com recurso às fórmulas do Excel. Posteriormente, na fase de passagem dos datasets para as dimensões que vão fazer parte do Data Warehouse foi utilizado o *Python*, que permitiu uma manipulação mais simples e veloz dos dados, tendo sido também utilizado para criar determinadas colunas de validação, dimensões sem tabela de origem e a juncão das fontes de dados e criação de chaves nas tabelas de factos.

Na última etapa do sistema ETL foi feito o carregamento dos dados extraídos e transformados para a *data presentation area*. Isto foi realizado através da criação de uma base de dados com recurso ao *PostgreSQL*, que serviu de certa forma também como uma última validação dos dados a partir de restrições de campos e chaves. Para esta fase utilizamos também o software *PowerBI*, que nos permite uma melhor visualização gráfica dos resultados dos relatórios.

O seguinte diagrama mostra o processo dos dados dentro do sistema ETL, desde a extração até ao carregamento nas ferramentas de análise.

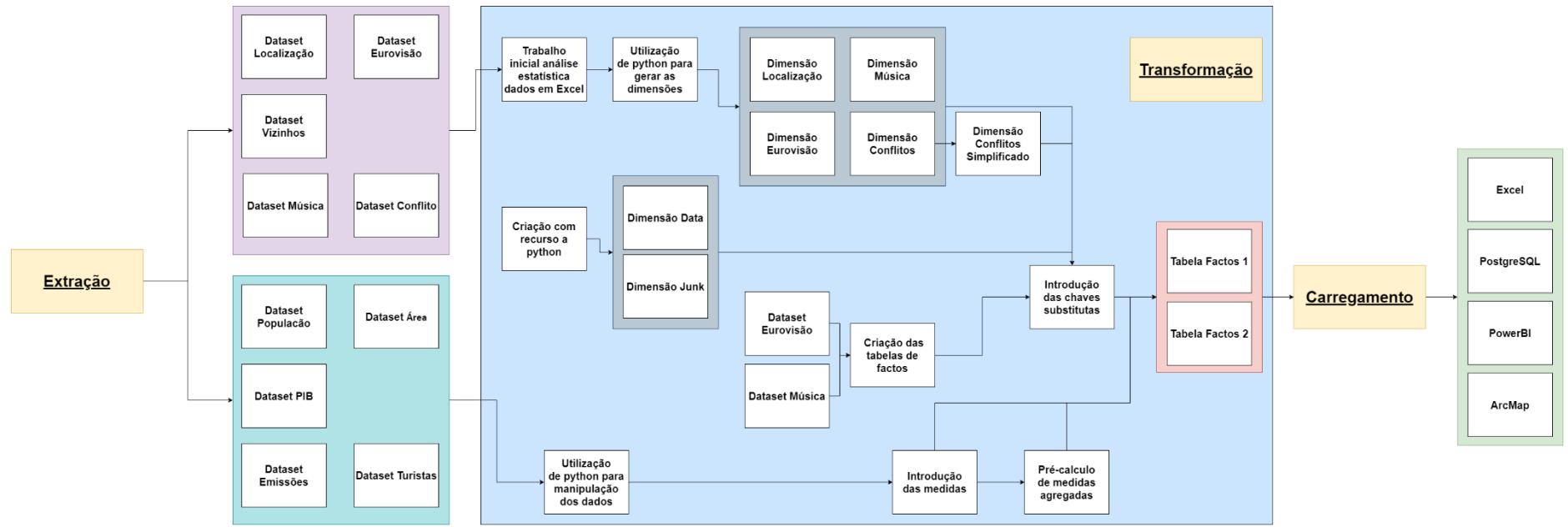


Figura 48 - Diagrama do sistema ETL

## 10. Dimensões e tabela de factos implementados no cubo de dados

Com vista gerar relatórios e poder responder às nossas questões analíticas utilizou-se o modelo no *Power BI*, visível na figura seguinte, necessitando, apenas, de realizar as ligações entre as dimensões e a tabela de factos através das chaves estrangeiras, à semelhança do esquema idealizado na etapa anterior.

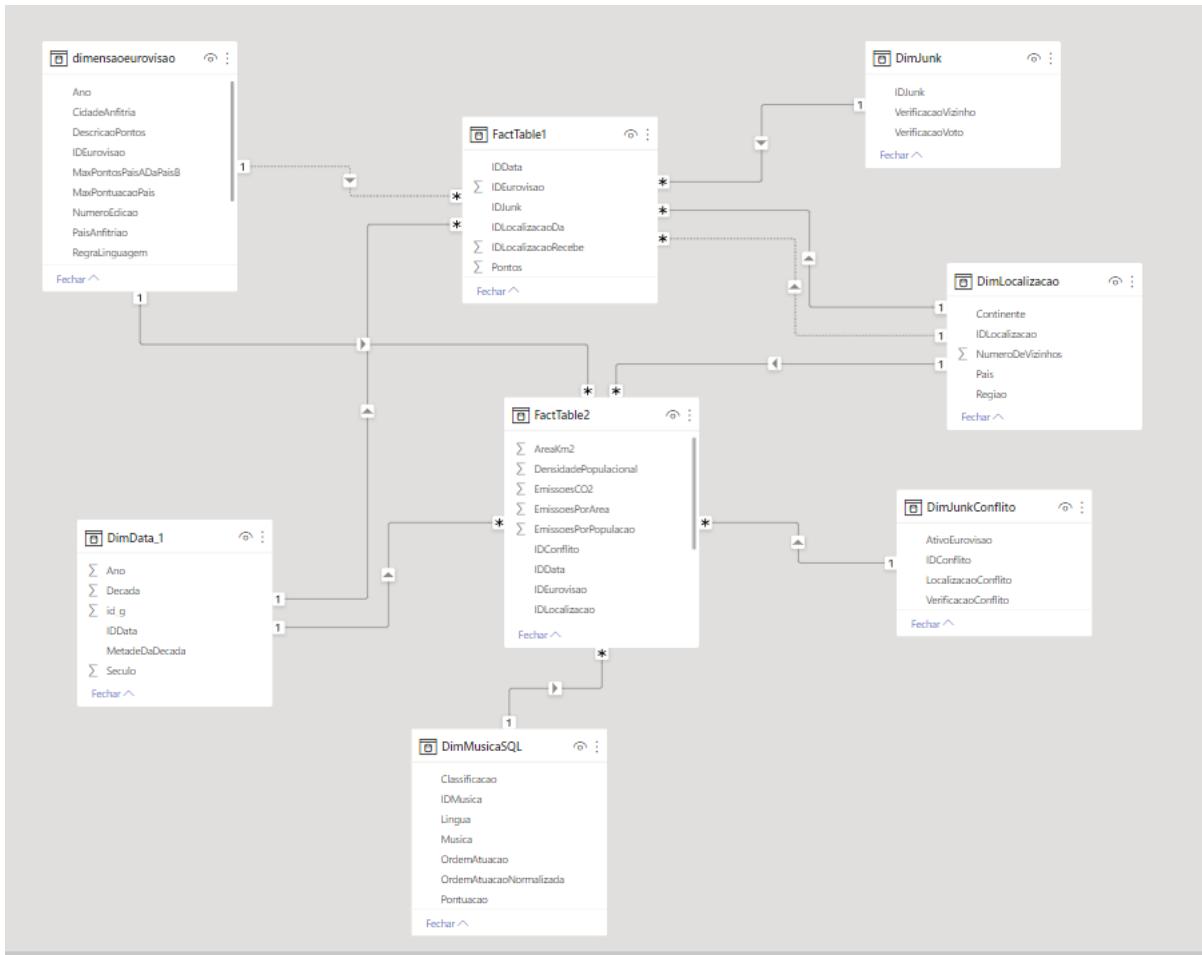


Figura 49: Implementação das dimensões e tabelas de factos utilizando o Power BI

Implementámos, ainda, as dimensões e as tabelas de factos no programa *postgreSQL*, visível nas figuras seguintes. Aqui é possível observar a estruturas das tabelas, estando representado um excerto das tabelas de dimensão na figura 50 e as tabelas de factos nas figuras 51 e 52.

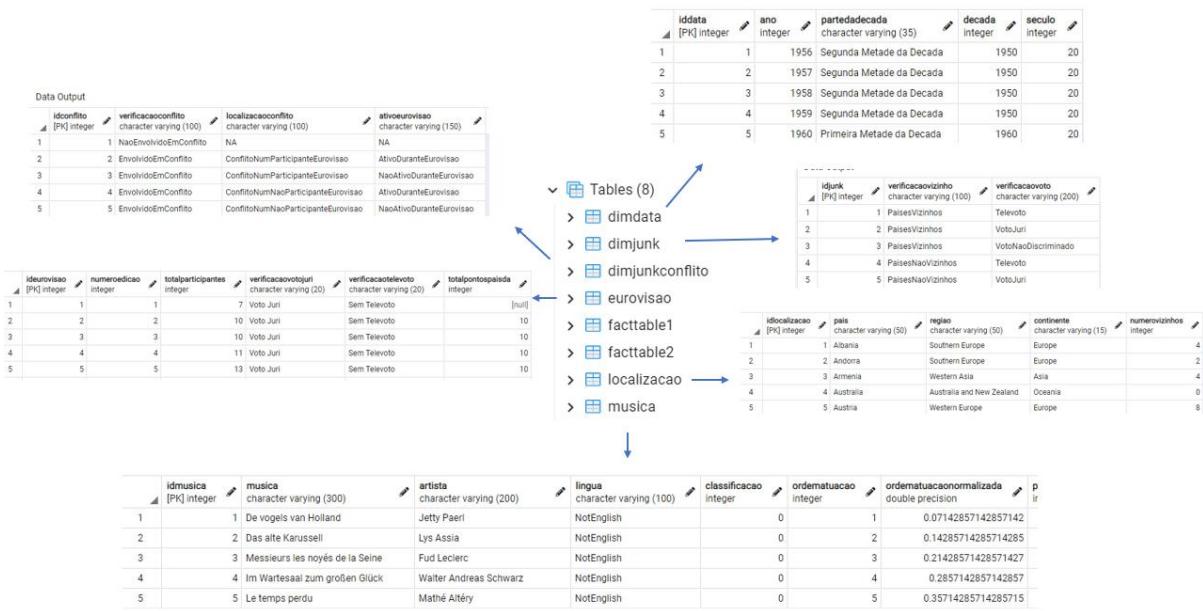


Figura 50: Implementação das dimensões e tabelas de factos utilizando o postgresSQL

	iddata [PK] integer	ideurovisao integer	idlocalizacaoda [PK] integer	idlocalizacaorecebe [PK] integer	idjunk [PK] integer	pontos integer
1		60	60	44	46	6
2		59	59	44	46	6
3		56	56	44	46	6
4		48	48	44	46	6
5		47	47	44	46	6
6		44	44	44	46	6
7		42	42	44	46	6
8		41	41	44	46	6
9		40	40	44	46	6
10		38	38	44	46	6
11		64	64	44	46	5
12		63	63	44	46	5
13		60	60	44	32	6
						10

Figura 51: Tabela de factos 1

	iddata integer	ideurovisao integer	idlocalizacao integer	idmusica [PK] integer	idconfiito integer	pontos integer	areakm2 double precision	pibpercapita double precision	populacao double precision	densidadepopulacional double precision	emissoesco2 double precision	emissoespord double precisi
69	49	49	36	1127	1	3	365244	57603.836021826	4591910	12.572170932308266	14.6	3.17950482
70	49	49	1	59	1	106	27400	2373.58129170055	3026939	110.4722627737226	3.18	1.05056626
71	49	49	2	77	1	0	470	37966.1872524395	76250	162.2340425391915	0.56	7.3442622
72	49	49	7	168	1	0	202810	2378.62328600741	9730146	47.976657955722104	30.99	3.18494707
73	49	49	9	252	1	91	51200	2698.4671799989	3764194	73.5194140625	15.92	4.2293256
74	49	49	11	287	1	50	55960	9744.04077059414	4304600	76.92280200142959	20.27	4.7089160
75	49	49	12	324	1	170	9240	23792.6213632243	1010410	109.3517316017316	7.45	7.3732445
76	49	49	15	406	1	0	42390	8914.10355674451	1362550	32.14319414956358	15.55	1.14124252
77	49	49	16	461	1	0	[null]	37702.8453762641	5228172	[null]	52.66	1.00723541
78	49	49	20	636	1	252	128900	21955.1041364941	10955141	84.98945694336695	98.76	9.0149456
79	49	49	22	687	1	16	100250	47334.9306537723	292074	2.9134563591022444	2.29	7.8404787
80	49	49	23	742	1	7	68890	47666.5470766131	4070262	59.083495427493105	42.73	1.04980956
81	49	49	24	786	2	0	21640	19910.6106710093	6809000	314.64879852125694	63.15	9.274485
82	49	49	25	820	1	0	60104	6770.666666677115	6667100	72.00000000000001	6.74	0.75702001

Figura 52: Tabela de factos 2

Uma vez que o nosso projeto não apresenta muitas medidas aditivas por isso a utilização do cubo de dados não faz muito sentido no âmbito deste projeto. No entanto realizamos algumas operações de cubo apenas para mostrar as capacidades da data *warehouse* montada no SQL, como o cálculo da soma dos pontos que um país recebe por ano, por verificação de vizinho ou não vizinho.

```

1 SELECT l.pais, j.verificacaovizinho, d.ano, SUM(ft.pontos)
2 FROM facttable1 ft, dimlocalizacao l, dimjunk j, dimdata d
3 WHERE ft.idlocalizacaorecebe=l.idlocalizacao AND j.idjunk=ft.idjunk AND d.iddata=ft.iddata
4 GROUP BY CUBE (l.pais, j.verificacaovizinho,d.Ano)
```

Figura 53 - Comando SQL para a criação de um cubo

	pais character varying (50) 	verificacaovizinho character varying (100) 	ano integer 	sum bigint 
1	Albania	PaisesNaoVizinhos	2004	107
2	Albania	PaisesNaoVizinhos	2005	18
3	Albania	PaisesNaoVizinhos	2008	10
4	Albania	PaisesNaoVizinhos	2009	17
5	Albania	PaisesNaoVizinhos	2010	20
6	Albania	PaisesNaoVizinhos	2012	44
7	Albania	PaisesNaoVizinhos	2015	6
8	Albania	PaisesNaoVizinhos	2018	75
9	Albania	PaisesNaoVizinhos	2019	44
10	Albania	PaisesNaoVizinhos	2021	32
11	Albania	PaisesNaoVizinhos	[null]	373
12	Albania	PaisesVizinhos	2004	28
13	Albania	PaisesVizinhos	2005	15
14	Albania	PaisesVizinhos	2008	10
15	Albania	PaisesVizinhos	2009	7
16	Albania	PaisesVizinhos	2010	10
17	Albania	PaisesVizinhos	2012	10
18	Albania	PaisesVizinhos	2015	16

Figura 54 - Exemplo de resultados do cubo

## 11. Produção de relatórios e resposta às questões analíticas

Para realizar os relatórios escolheram-se os programas *Power BI* e *PostgreSQL*. Esta escolha deveu-se ao facto do primeiro ser uma ferramenta utilizada para *Business Intelligence*, permitindo realizar análises e tirar conclusões de uma forma rápida. Já o segundo programa consiste em utilizar a linguagem SQL - que é um fator positivo - pois esta linguagem é de extrema relevância quando o assunto consiste em manipular os dados. Esta é ideal quando queremos trabalhar com grandes volumes de dados de forma rápida e segura.

### 11.1 Primeira pergunta analítica – Influência da língua da canção

*"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?"*

Para conseguirmos obter uma resposta satisfatória a esta pergunta, procurámos primeiro responder a cada uma das sub-perguntas que a compõem.

#### 11.1.1 Quantidade de países que não se qualificam para a final – em inglês e língua materna

Para responder a esta pergunta, recorreu-se ao programa *postgreSQL*, utilizando código SQL.

O seguinte código consiste em determinar a razão entre o número de músicas que não foram qualificadas para a final cuja língua não era o inglês com o número total de músicas (qualificadas e não qualificadas) cuja língua era o inglês, visível na figura 55. Realizou-se o mesmo procedimento para as músicas cuja língua era o inglês e a junção do inglês com a língua materna, figuras 56 e 57 respetivamente.

Dado que só começaram a existir semifinais em 2004, determinou-se o impacto da língua a partir dessa data.

```
1  SELECT
2    (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
3     FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
4    WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA AND
5      ORDEMATUACAONORMALIZADA IS NULL AND LINGUA = 'NotEnglish' AND ANO>=2004 )/
6    (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
7     FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
8    WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA AND
9      LINGUA = 'NotEnglish' AND ANO>=2004 ) AS RESULTADO
```

Data Output    Messages    Notifications    Explain

	resultado
1	double precision
1	0.375

Figura 55: Código SQL utilizado e respetivo resultado obtido para a língua Não Inglês

```

1 SELECT
2 (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
3 FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
4 WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA and
5 ORDEMATUACAONORMALIZADA IS NULL AND LINGUA = 'English' AND ANO>=2004 )/
6 (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
7 FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
8 WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA AND
9 LINGUA = 'English' AND ANO>=2004 ) AS RESULTADO

```

Data Output Messages Notifications Explain

	resultado	
	double precision	lock
1	0.32108317214700194	

Figura 56: Código SQL utilizado e respetivo resultado obtido para a língua inglês

```

1 SELECT
2 (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
3 FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
4 WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA and
5 ORDEMATUACAONORMALIZADA IS NULL AND LINGUA = 'Mixed' AND ANO>=2004 )/
6 (SELECT CAST(COUNT(t.idmusica) AS FLOAT)
7 FROM PUBLIC.MUSICA m, Public.facttable2 t, Public.dimdata d
8 WHERE m.IDMUSICA=t.IDMUSICA AND d.IDDATA=t.IDDATA AND
9 LINGUA = 'Mixed' AND ANO>=2004 ) AS RESULTADO

```

Data Output Messages Notifications Explain

	resultado	
	double precision	lock
1	0.3333333333333333	

Figura 57: Código SQL utilizado e respetivo resultado obtido para a língua Mixed

Com base nas três figuras anteriores, verifica-se que, a partir de 2004, as músicas que foram cantadas na língua materna do país, apresentam menos qualificações para a final uma vez que o resultado obtido é 0.37 para não inglês enquanto para inglês e Mixed é 0.32 e 0.33 respetivamente.

### 11.1.2 As músicas em inglês obtêm melhores resultados?

Para responder a esta pergunta *recorreu-se ao programa Power BI*, através da utilização da visualização do gráfico de colunas empilhadas.

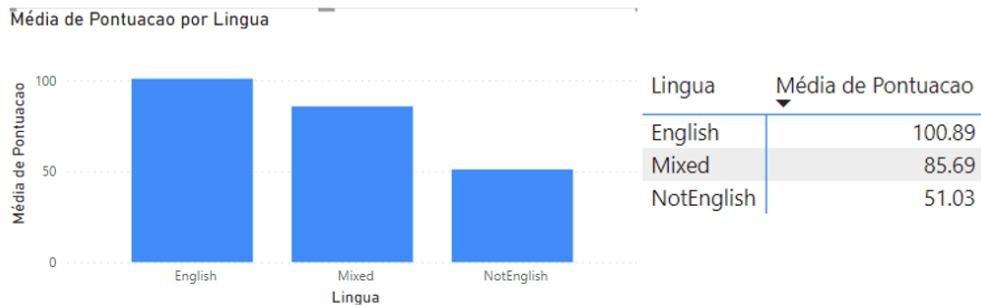


Figura 58: Média das pontuações por tipo de linguagem da música

Ao observar a figura acima, é visível que as músicas que são cantadas na língua materna do país apresentam pior pontuação quando comparadas com as músicas cantas só em inglês e com a junção do inglês com a língua materna.

No entanto, ao longo do festival as regras foram mudando e a certa altura eliminou-se a regra da linguagem – consistia em cada país ter de cantar obrigatoriamente com a sua língua materna.

Para determinar o efeito do antes e depois da regra da linguagem, realizaram-se os seguintes gráficos de colunas empilhadas.

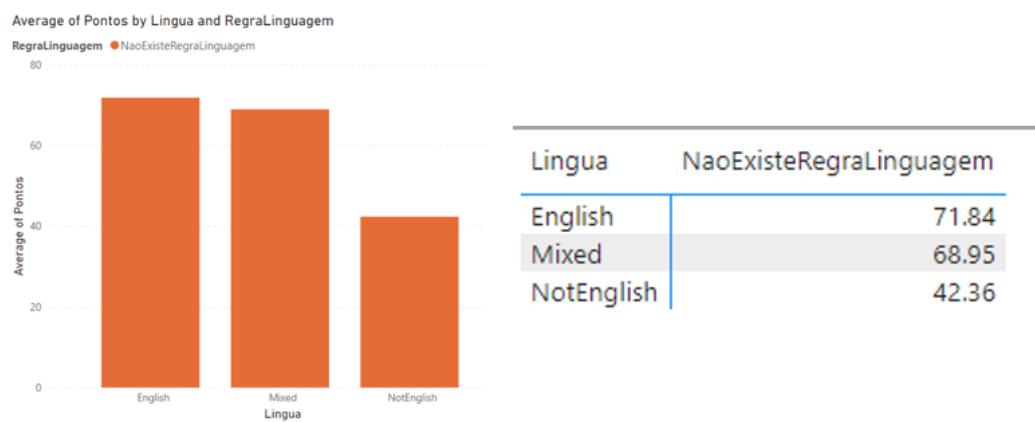


Figura 59: Média de pontos quando não existe regra de linguagem

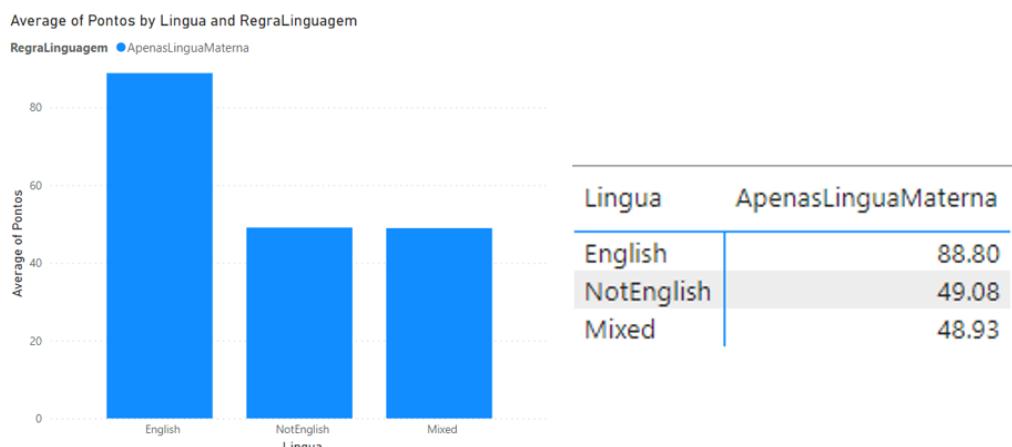


Figura 60: Média de pontos quando existe regra de linguagem

Ao observarmos as figuras 59 e 60, conclui-se que independentemente de existir ou não existir regra de linguagem, a música em inglês obtém no geral mais pontos.

No entanto é interessante observar a influência na classificação.

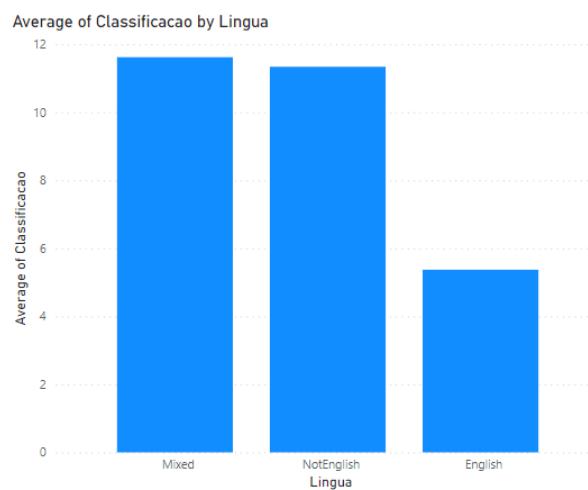


Figura 61: Classificação média por língua quando existe regra de língua materna

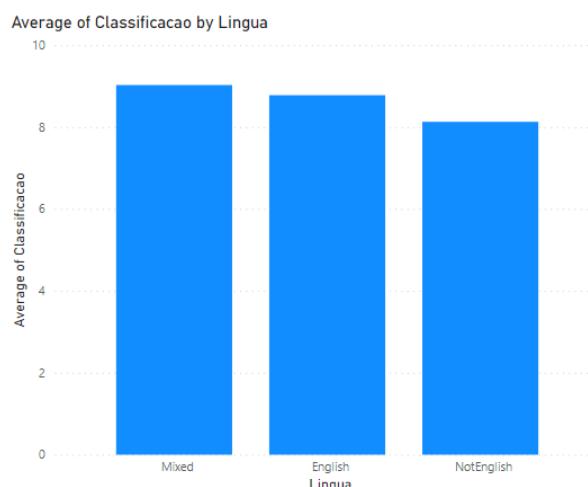


Figura 62: Classificação média por língua quando não existe regra de língua materna

Ao observar a figura 61, é visível que quando existia regra de linguagem, as músicas em inglês ficavam mais bem classificadas, no entanto, atualmente, quando já não existe regra de linguagem – figura 62 - as músicas que são cantadas na língua materna dos participantes obtém melhores classificações.

Uma vez que a dimensão data continha uma hierarquia na qual se encontrava o campo década, decidiu-se analisar a influencia da língua ao longo de cada década, visível na figura seguinte.

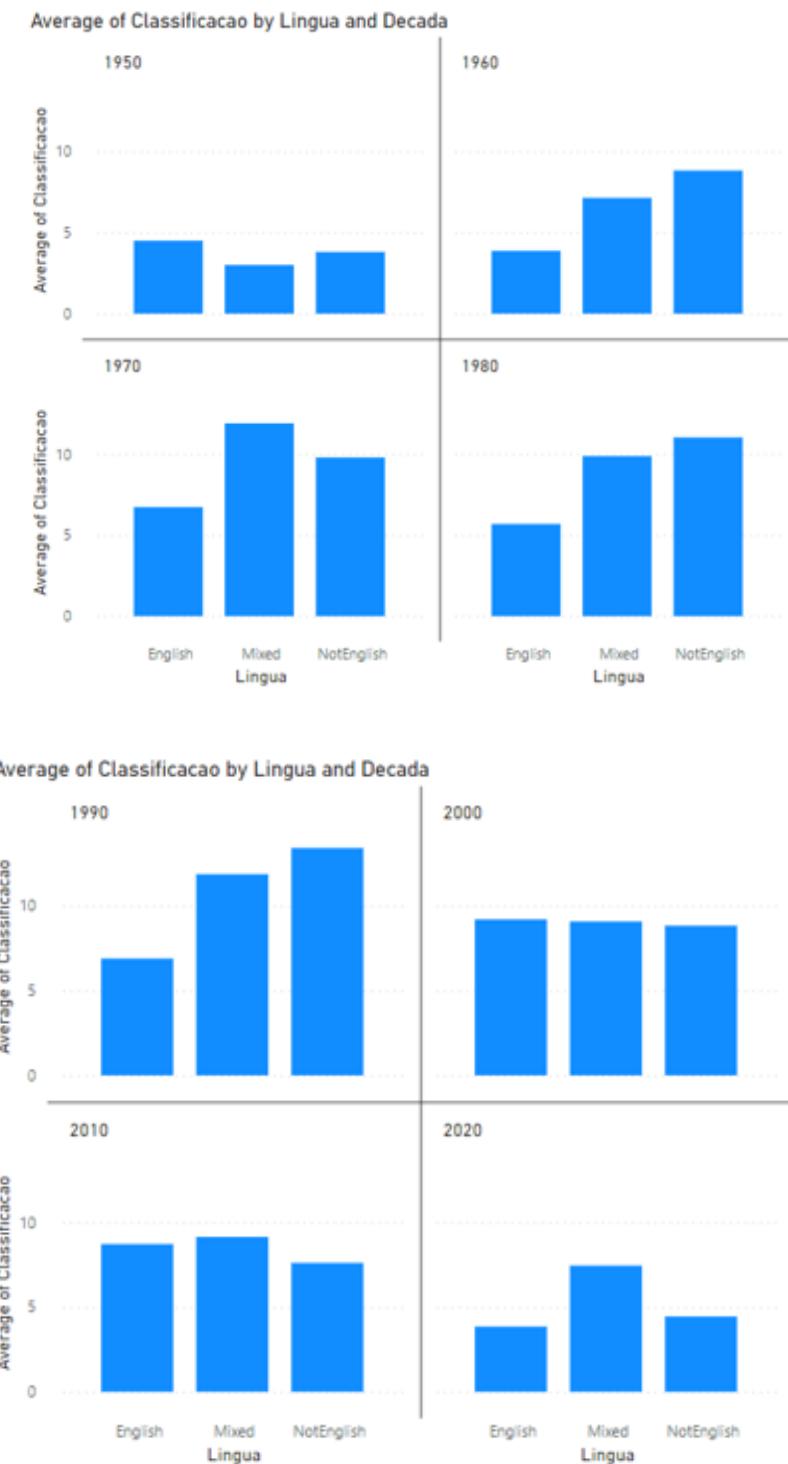


Figura 63: Média das classificações por década

Na figura 63 é visível que até à década de 1990, as músicas cuja linguagem é o inglês apresentavam claramente os melhores resultados, no entanto desde os anos 2000 que a tendência tem sido invertida, sendo que as músicas cantadas em línguas maternas têm vindo a ganhar popularidade.

Finalmente, analisamos a prevalência da língua nas músicas que terminaram em primeiro lugar e no top 5 do concurso.

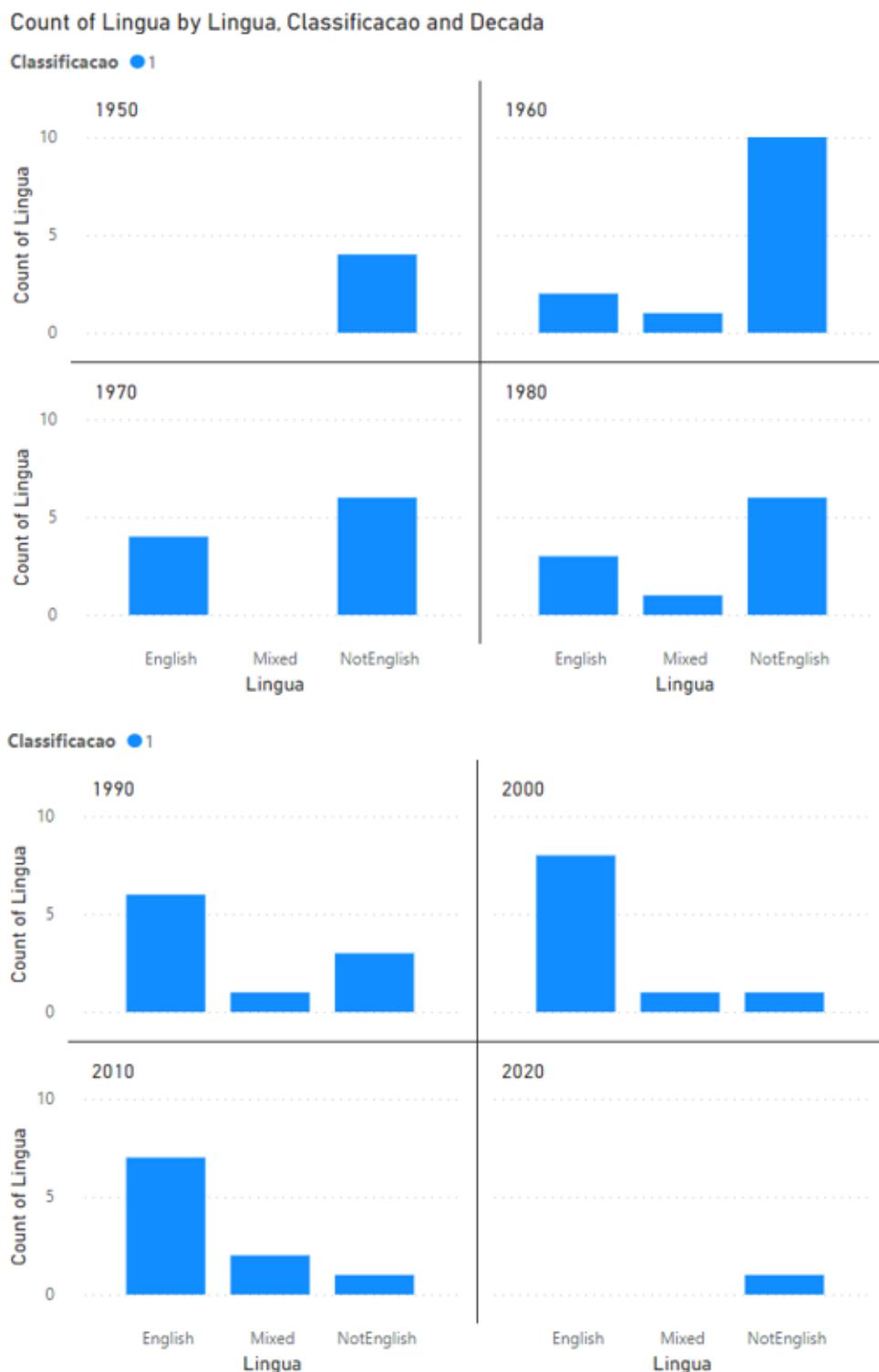


Figura 64 - Primeiros classificados por língua ao longo das décadas

Pela análise das figuras acima é visível que até à década de 80, sendo que na década de 90 as músicas em inglês começaram a ganhar mais frequentemente o concurso. Nas décadas de 2000 e 2010 apenas uma música cantada em língua materna ganhou o concurso. Sendo que na década de 2020 existem apenas dados para o ano de 2021 não existem ainda dados suficientes que nos permitam dizer se a tendência está ou não em mudança.

Count of Lingua by Lingua, Classificacao and Decada

Classificacao ● 1 ● 2 ● 3 ● 4 ● 5

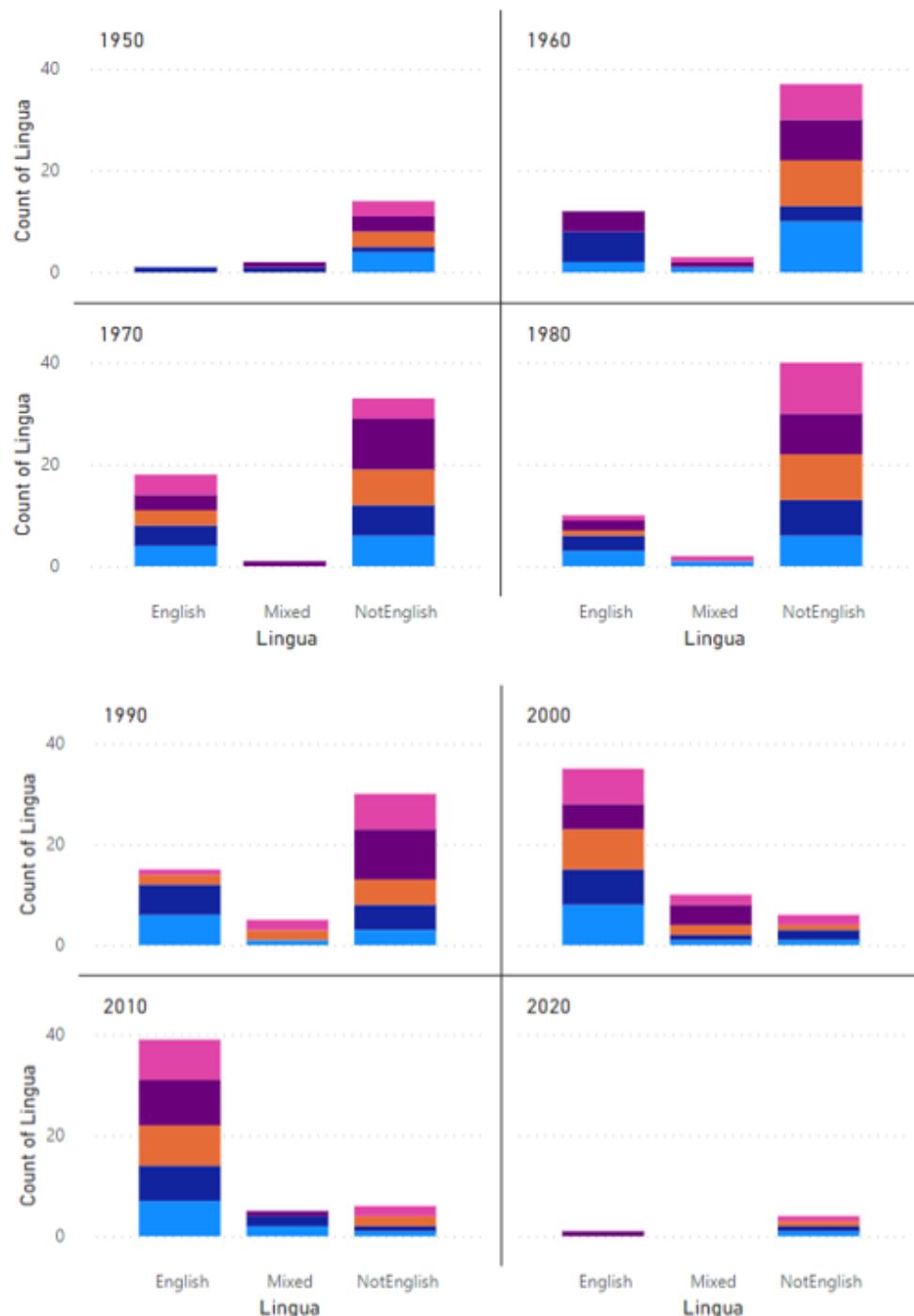


Figura 65 - Top 5 classificados por língua ao longo das décadas

Para os resultados dos cinco primeiros classificados existiu também uma mudança abrupta, mas neste caso na década de 2000, para a preferência das músicas em inglês. Mais uma vez, a década de 2020 apresenta um melhor resultado para as músicas em língua materna, mas ainda não existem dados suficientes para concluir se existiu uma mudança de preferência.

### 11.1.3 Diferença dos resultados de um país quando canta em inglês versus língua materna

Para responder a esta pergunta foram utilizados os dois programas – PostgreSQL e Power BI.

O primeiro consistiu em determinar a média dos pontos obtidos por cada país pelo tipo de língua - estando alguns valores a Null, o que significa que o país nunca cantou nessa língua, visível na figura 66.

Query Editor Query History			
Data Output Messages Notifications Explain			
	avg double precision	pais character varying (200)	lingua character varying (100)
1	12.33333333333334	Albania	English
2	14	Albania	NotEnglish
3	[null]	Albania	Mixed
4	14	Andorra	NotEnglish
5	[null]	Andorra	Mixed
6	12.375	Armenia	English
7	[null]	Armenia	NotEnglish
8	7.33333333333333	Armenia	Mixed

Figura 66: Código SQL utilizado e respetivos resultados obtidos

Para se conseguir obter uma melhor percepção dos resultados obtidos, importou-se a tabela resultante das *queries* para o Power BI e visualizou-se os resultados através da visualização de um gráfico de colunas agrupadas. Neste é possível verificar se o país obteve melhores ou piores resultados utilizando a língua inglesa como língua oficial da música.

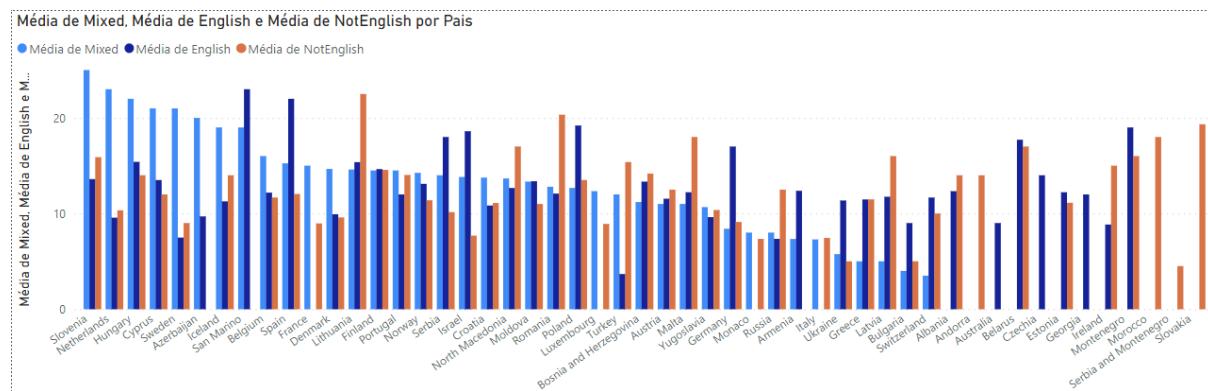


Figura 67: Comparação das médias das classificações obtidas por cada país por língua

Na figura 67, observa-se que a maior parte dos países apresentaram músicas tanto em inglês, na sua língua materna e com a mistura das duas, no entanto, não é possível verificar uma relação direta entre cantar em inglês e obter melhor resultados. Contudo, verifica-se que há determinados grupos de países que apresentam melhores classificações cantando as suas músicas em inglês e outro grupo de países que apresentam melhores classificações cantando as músicas na sua língua materna.

Com o intuito de perceber melhor o gráfico anterior, foi realizada uma subtração entre as médias de classificação para os vários critérios da língua, eliminando os países que nunca cantaram em inglês ou nunca se qualificaram para a final com uma música em inglês e os países que apenas cantaram em inglês, fez-se a diferença entre os resultados obtidos quando cantaram em inglês *versus* quando cantaram em língua materna.

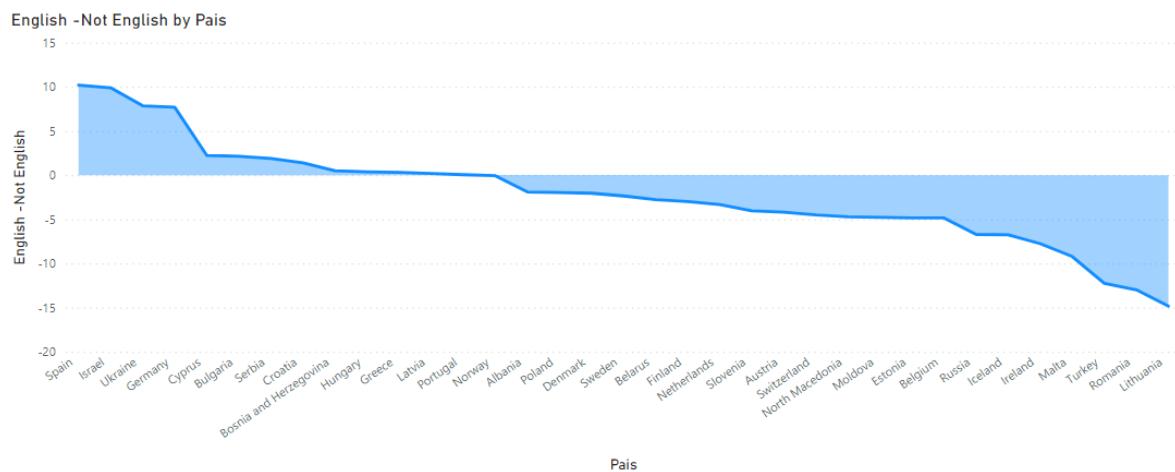


Figura 68: Comparação de resultados obtidos pelo mesmo país cantando músicas em inglês e não inglês.

Na figura 68 é visível que os países à esquerda apresentam melhores resultados em língua materna enquanto os resultados à direita apresentam melhores resultados em inglês. No geral, os resultados são melhores quando os países cantam em inglês.

Fazendo a mesma análise para as músicas em inglês *versus* as músicas cantadas em mistura de línguas – Figura 69 - os resultados no geral são mais balançados, mas ainda assim existe uma leve tendência para melhores resultados em inglês. Também nesta figura os resultados à direita apresentam melhores resultados em inglês e à esquerda melhores resultados em mistura de línguas.

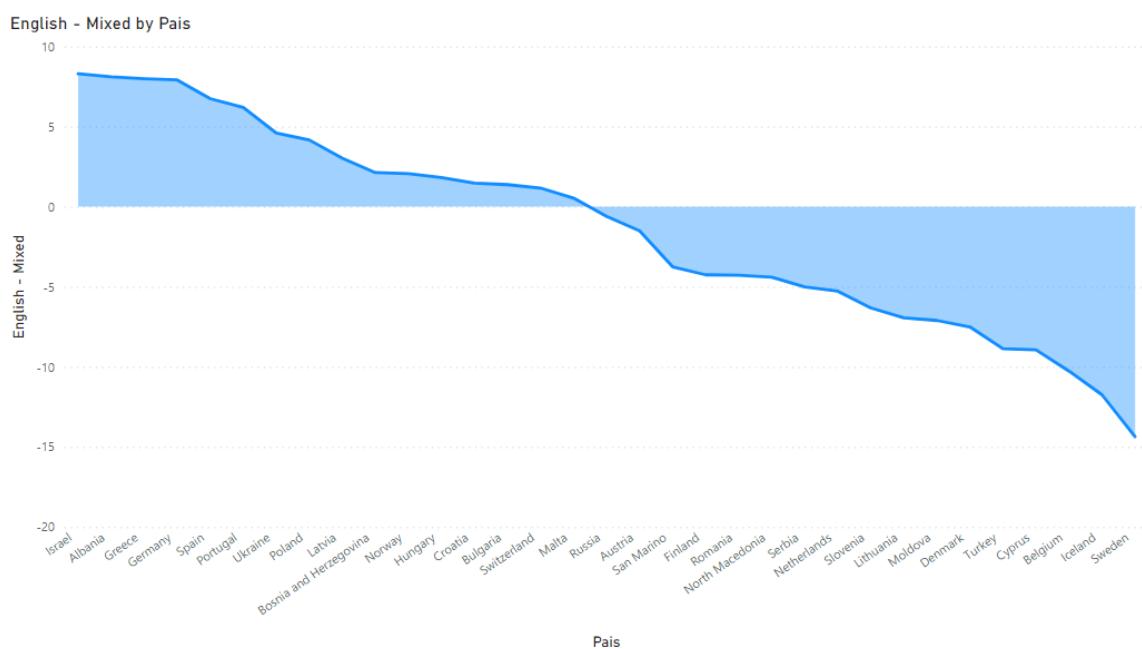


Figura 69: Comparação de resultados obtidos pelo mesmo país cantando músicas em inglês e mistura do inglês com a sua língua materna

## 11.2 Segunda pergunta analítica – Influência da demografia e geografia

*“Como é que a demografia e a geografia influenciam os resultados na eurovisão? O número de vizinhos de um país tem influência na quantidade de pontos que este recebe? Existe entreajuda entre vizinhos? Os países com maior população vizinha têm vantagens? Os países com maior PIB têm melhores resultados?”*

Mais uma vez, para responder a esta pergunta procurámos primeiro responder às sub-perguntas, descritas em seguida.

### 11.2.1 Influência do número de vizinhos de um país na quantidade de pontos recebidos

Para responder a esta pergunta, recorreu-se ao programa *PowerBI* e foi utilizada a visualização em gráfico de linhas e colunas empilhadas, visível na figura seguinte. No eixo do XX, encontram-se os países que participam na eurovisão, já no eixo dos YY, encontra-se o número de vizinhos que um país tem. A reta que se encontra na figura corresponde ao valor média da classificação de cada país.

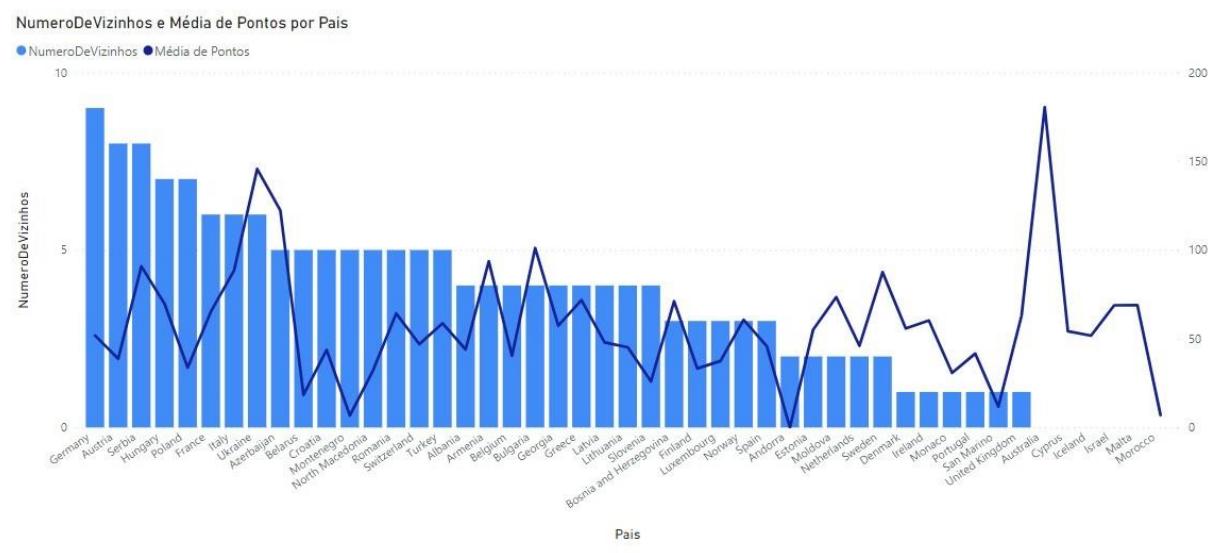


Figura 70: Número de vizinhos que cada país tem e a respetiva média de pontos que obtém no festival.

Na figura 70 conseguimos verificar que não existe uma relação direta entre o número de países e a média de pontos recebidos, no entanto esta análise ao nível do país não é suficiente para conseguir obter conclusões do panorama geral.

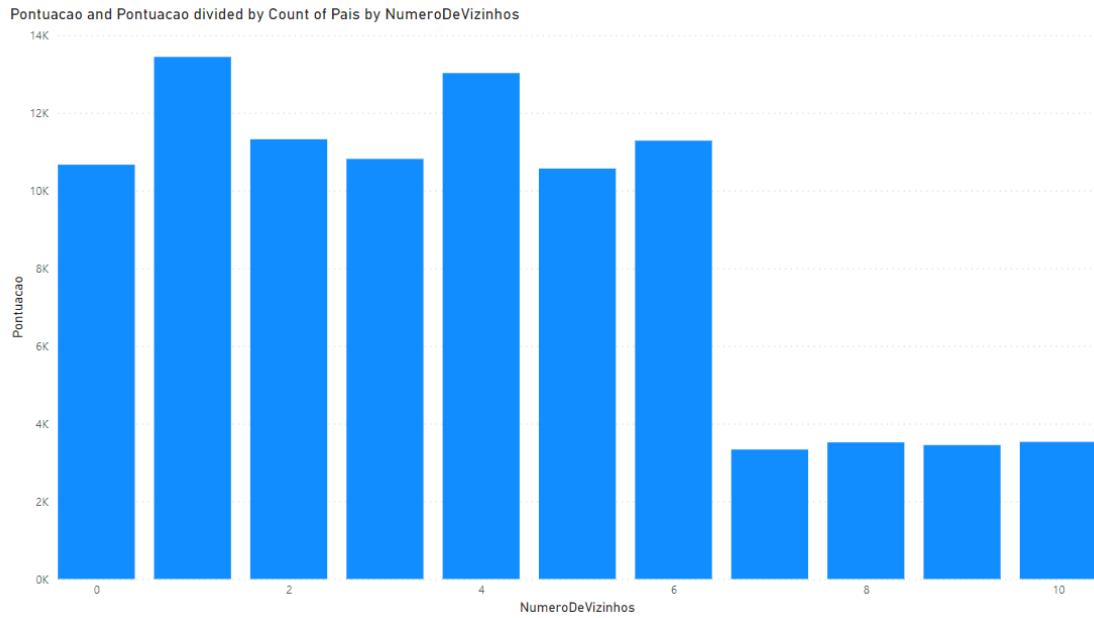


Figura 71: Soma da pontuação total que cada país dá ao seu vizinho.

Ao observarmos a figura 71, deparamo-nos com o problema de que há não há o mesmo número de países com o mesmo número de vizinhos, como tal, sentimos a necessidade de normalizar.

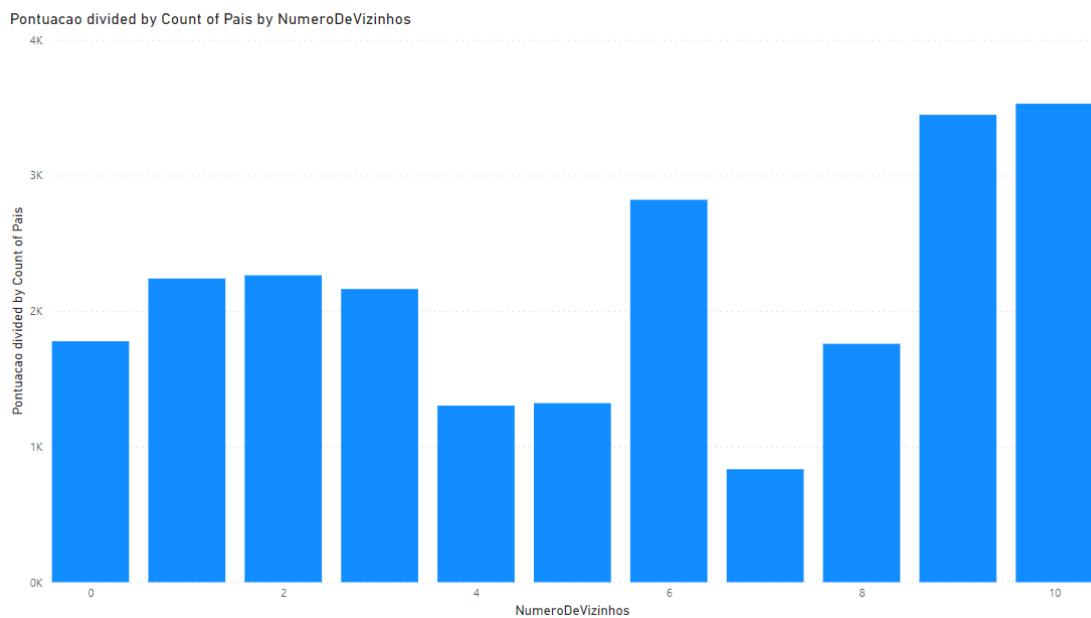


Figura 72: Soma de pontos dividida por número de países com número x de vizinhos versus número de vizinhos

Ao observarmos a figura 72, verificamos que os países com mais vizinhos – 9 e 10 vizinhos – são os que recebem mais pontos, no entanto há alguns outliers como é o caso dos países que apresentam 8 e 7 vizinhos.

À semelhança do caso anterior, decidimos observar se o número de vizinhos apresenta influência na classificação.

Average of Classificacao by NumeroDeVizinhos

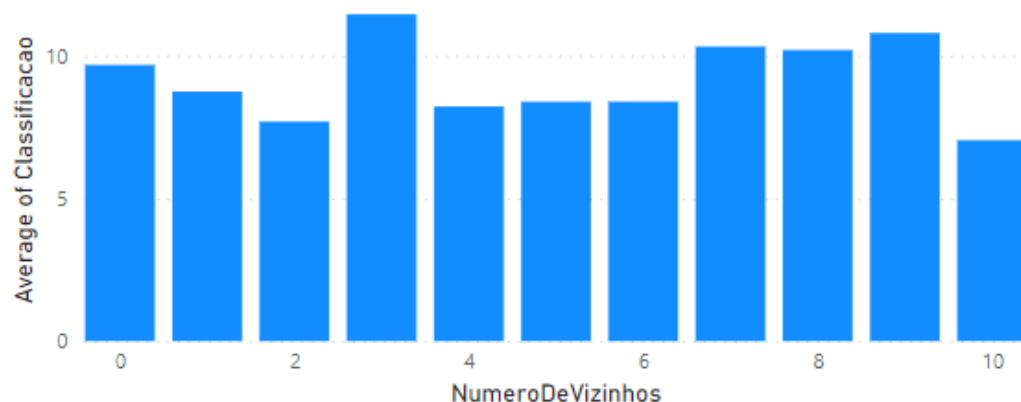


Figura 73: Classificação média dos países de acordo com o seu número de vizinhos.

Classificacao by NumeroDeVizinhos



Figura 74: Soma da Classificação por número de vizinhos.

Mais uma vez, foi necessário normalizar os resultados, pois não existe o mesmo número de países para o mesmo número de vizinhos.

Classificacao divided by Count of Pais by NumeroDeVizinhos

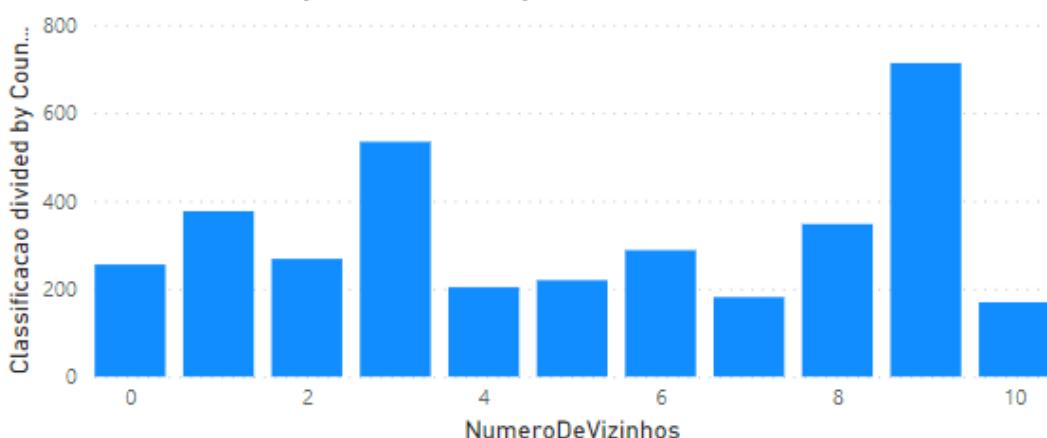


Figura 75: Classificação por número de vizinhos dividido pelo número de países com x vizinhos

Ao observarmos a figura 75, verifica-se que o país com 10 vizinhos apresenta melhores resultados no entanto o país com 9 vizinhos apresenta piores classificações quando comparado com os países com 8 vizinhos. Como existe apenas um país com 9 vizinhos, a Alemanha, este resultado não quer necessariamente dizer que não existe relação entre o número de vizinhos e o resultado final, podendo tratar-se de um outlier. Os países que tem entre 4 a 8 vizinhos apresentam no geral classificações melhores ou iguais aos países com menores números de vizinhos.

De seguida, fomos procurar qual o vizinho mais valioso, que consiste no vizinho que geralmente dá mais pontos dos seus países vizinhos e também o país que mais pontos recebe dos seus vizinhos por edição. Para tal, recorreu-se primeiro a interrogações no programa *PostgreSQL*.

```

1 SELECT SUM(f.Pontos) AS TotalPontos, l.Pais
2 FROM facttable1 f, dimlocalizacao l, dimjunk j
3 WHERE f.IDLocalizacaoDa= l.IDLocalizacao AND j.IDjunk=f.IDjunk AND j.VerificacaoVizinho = 'PaisesVizinhos'
4 GROUP BY l.pais
5 ORDER BY TotalPontos DESC

```

Figura 76 - Comandos SQL para procura do vizinho mais valioso

	totalPontos	pais
	bigint	character varying (50)
1	951	Germany
2	620	France
3	529	Switzerland
4	524	Belgium
5	500	Norway
6	485	Russia
7	430	Austria
8	373	Sweden
9	368	Finland
10	356	Spain
11	342	Italy
12	307	Ukraine
13	304	Netherlands

Figura 77 - Resultado obtido

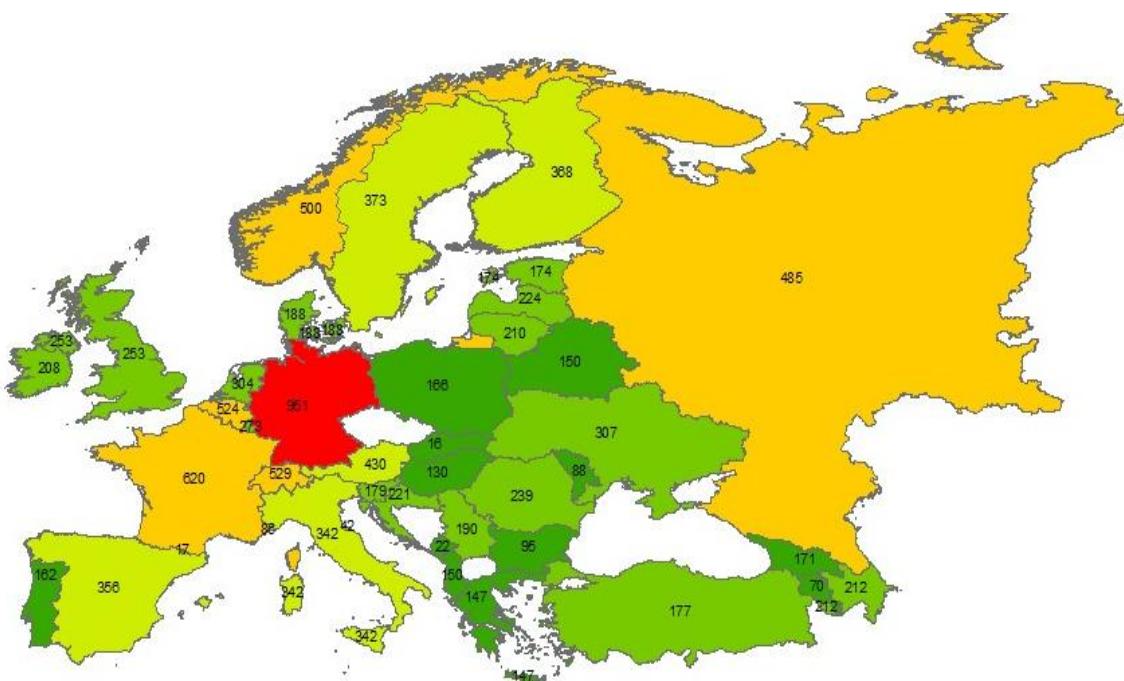
```

1 SELECT SUM(f.Pontos) AS TotalPontos, l.Pais, l.numerovizinhos
2 FROM facttable1 f, dimlocalizacao l, dimjunk j
3 WHERE f.IDLocalizacaoRecebe= l.IDLocalizacao AND j.IDjunk=f.IDjunk AND j.VerificacaoVizinho = 'PaisesVizinhos'
4 GROUP BY l.pais,l.numerovizinhos
5 ORDER BY TotalPontos DESC

```

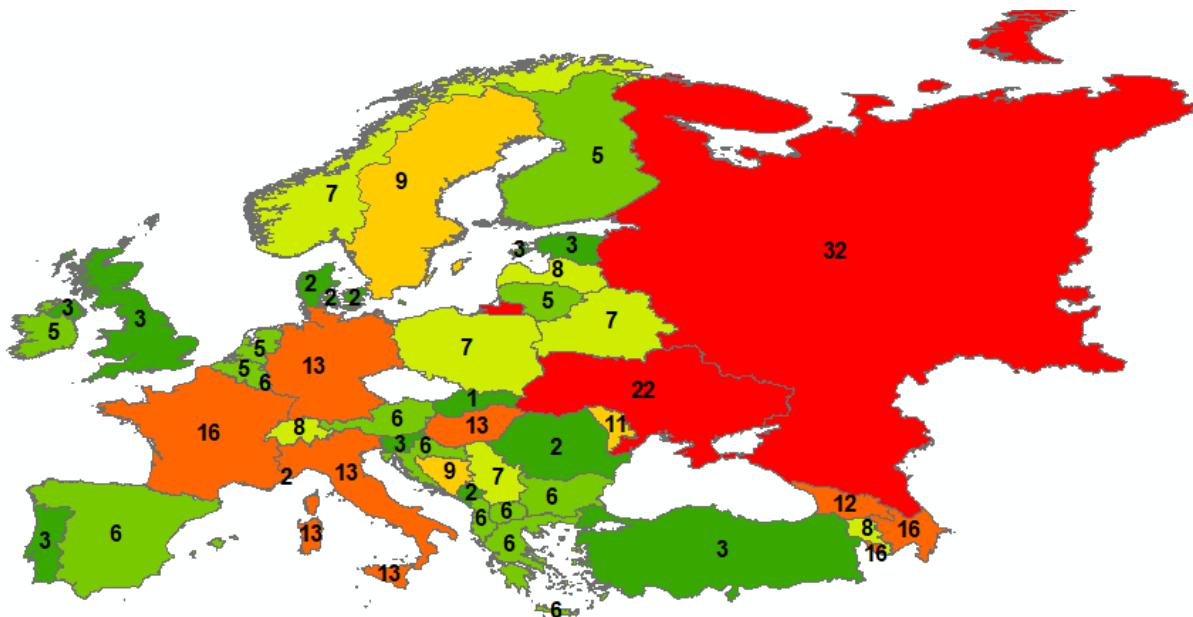
Figura 78: Código SQL utilizado para procura do país que mais pontos recebe dos seus vizinhos

Em seguida, os resultados foram exportados para o *ArcMap*. Foi em primeiro lugar obtido um ficheiro *ESRI Shapefile* com o mapa da Europa e o nome dos países e importado para o software as tabelas resultantes das interrogações realizadas em SQL e a dimensão localização. As tabelas foram depois ligadas através de operações de juncção e os resultados obtidos encontram-se nas figuras 77 e 78. Foram removidos os países sem vizinhos para facilitar a visualização.



*Figura 79: Mapa de pontos dados aos países vizinhos em todas as edições da Eurovisão*

Na figura 79 podemos observar que o vizinho mais valioso é a Alemanha, que ao longo dos anos tem dado um maior número de pontos aos seus países vizinhos. Por outro lado, os países que menos pontos deram aos seus vizinhos foram Montenegro e Eslováquia.



*Figura 80 - Média de pontos que cada país recebe por edição pelos seus vizinhos.*

Na figura 80 podemos ver a média de pontos que cada país recebe por edição dos seus vizinhos. A Rússia e a Ucrânia são os países que mais pontos recebem dos seus vizinhos, sendo que estes países têm 10 e 6 vizinhos que participam na Eurovisão, respectivamente.

### 11.2.2 Entreajuda entre vizinhos ou outros países

Para analisar se existem pares de vizinhos ou não vizinhos que se repetem com grande frequência nos festivais foi realizada uma análise de combinações de países que mais se repetem, através de uma regra de associação simples.

Inicialmente houve uma tentativa de implementar o código no *Python* mas devido ao grande tamanho dos dados este não estava a conseguir realizar operações, pelo que se optou por fazer uma operação simples de *Count IF* no Excel e divisão pelo numero de “cestos” - numero de edições da Eurovisão, 65 - para obter o suporte de cada combinação de países.

De seguida, apresentamos um excerto da tabela resultante, onde são visíveis os resultados para os países que mais vezes votam num país, sendo que os resultados seguintes são todos de países que votaram mais de 50% das edições num outro país.

A distribuição entre países vizinhos e não vizinhos é aproximadamente igual. Devido às limitações do Excel não foi possível analisar estas combinações num segundo nível (seria, por exemplo, interessante ver o suporte de pares de países que dão pontos entre si – ex. *Support* ({NorwaySweden, SwedenNorway}).

Tabela 76- Combinações mais populares de países

Combinações de Países	Count	Suporte	Vizinho
<b>Norway - Sweden</b>	46	0.707692	Sim
<b>Sweden -Norway</b>	39	0.6	Sim
<b>Ireland -United Kingdom</b>	38	0.584615	Sim
<b>United Kingdom - Sweden</b>	38	0.584615	Não
<b>Sweden - France</b>	37	0.569231	Não
<b>Denmark - Sweden</b>	36	0.553846	Não
<b>Spain - Italy</b>	36	0.553846	Não
<b>Switzerland -United Kingdom</b>	36	0.553846	Não
<b>United Kingdom - Ireland</b>	36	0.553846	Sim
<b>France - Sweden</b>	35	0.538462	Não
<b>Germany - United Kingdom</b>	35	0.538462	Não
<b>Spain - Germany</b>	35	0.538462	Não
<b>Finland - Sweden</b>	34	0.523077	Sim
<b>Norway - France</b>	34	0.523077	Não
<b>Sweden - Denmark</b>	34	0.523077	Não
<b>Switzerland - France</b>	34	0.523077	Sim
<b>Germany - France</b>	33	0.507692	Sim
<b>Spain - Sweden</b>	33	0.507692	Não
<b>Sweden - Ireland</b>	33	0.507692	Não
<b>United Kingdom - Germany</b>	33	0.507692	Não

Ainda assim, são visíveis nesta tabela cinco pares de países em que tanto o país A a dar pontos ao país B como o país B a dar pontos ao país A tem suporte acima de 0.5. Esses pares são (Noruega,Suécia), (Irlanda,Reino Unido), (França, Suécia), (Suécia,Dinamarca) e (Reino Unido, Alemanha). Em dois destes casos os países são vizinhos e em outros dois casos os países são geograficamente próximos.

### 11.2.3 Influência da população vizinha dos países

Para responder a esta pergunta recorreu-se ao programa *Power BI* e utilizando a visualização Gráfico de linhas e colunas empilhadas, visível na figura seguinte. No eixo do XX, encontram-se os países que participam na eurovisão, já no eixo dos YY, encontra-se os valores da População. A reta que se encontra na figura corresponde ao valor média da classificação de cada país.

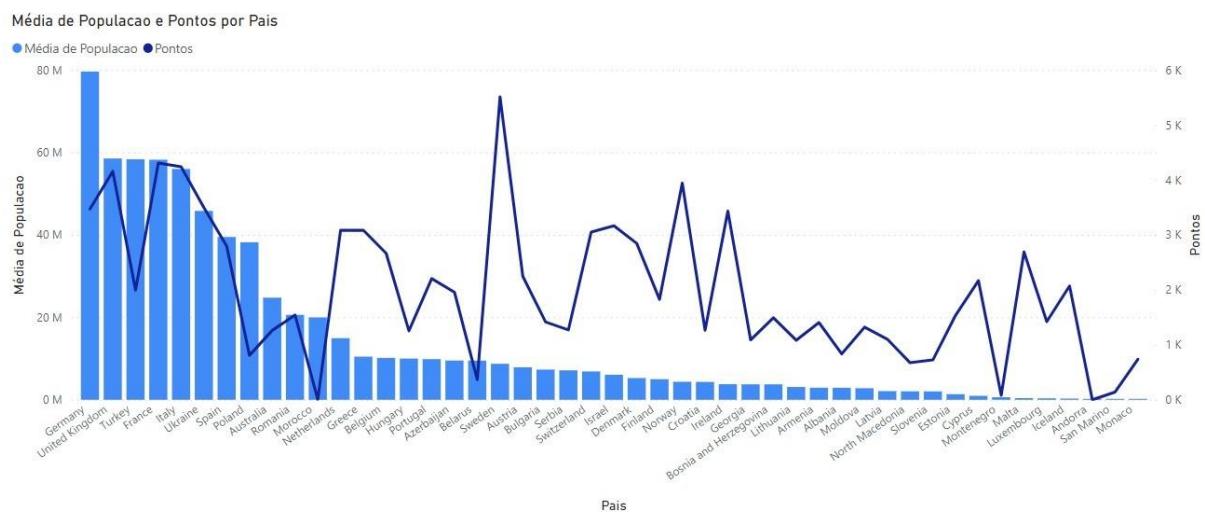


Figura 81: Comparação das médias das classificações obtidas por cada país com a população de cada país.

Ao observar a figura 81, conseguimos perceber que não há correlação entre a População de um país com a sua classificação no festival

### 11.2.4 Os países com maior PIB têm melhores resultados?

Para responder a esta pergunta recorreu-se ao programa *Power BI* e utilizando a visualização Gráfico de linhas e colunas empilhadas, visível na figura seguinte. No eixo do XX, encontram-se os países que participam na eurovisão, já no eixo dos YY, encontra-se os valores do PIB. A reta que se encontra na figura corresponde ao valor média da classificação de cada país.

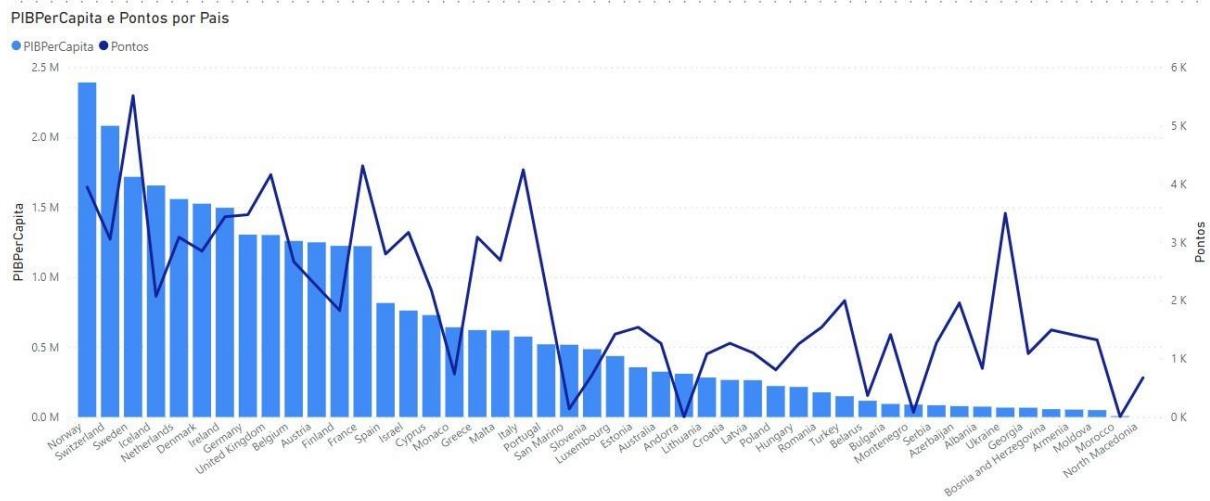


Figura 82: Comparação das médias das classificações obtidas por cada país com o PIBPerCapita de cada país.

Ao observar a figura 82, conseguimos perceber que não há correlação do PIBPerCapita com a classificação dos países no festival.

### 11.3 Terceira pergunta analítica - Influência das questões da atualidade

*“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?”*

#### 11.3.1 A participação em conflitos diminui a média de pontos que um país recebe?

Para responder a esta pergunta utilizou-se os dois programas – PostgreSQL e Power BI. O primeiro foi utilizado para determinar as médias das classificações dos países quando não estavam em conflitos – figura 83 – e quando estavam em conflitos – figura 84.

```

1 SELECT AVG(classificacao)AS CLASSIFICACAO, l.pais
2 FROM Public.musica m, Public.facttable2 f, Public.localizacao l, Public.dimjunkconflito j
3 WHERE m.idmusica=f.idmusica AND l.idlocalizacao=f.idlocalizacao AND j.idconflito=f.idconflito
4 AND f.idconflito = 1
5 GROUP By pais

```

	Data Output	Messages	Notifications	Explain
	classificacao	pais		
1	14	Andorra		
2	12.9090909090908	Cyprus		
3	12.4	Turkey		
4	9.941176470588236	Switzerland		
5	7.166666666666667	Italy		
6	15.071428571428571	Hungary		

Figura 83: Código SQL utilizado e respetivos resultados obtidos

```

1 SELECT AVG(classificacao)AS CLASSIFICACAO, l.pais
2 FROM Public.musica m, Public.facttable2 f, Public.localizacao l, Public.dimjunkconflito j
3 WHERE m.idmusica=f.idmusica AND l.idlocalizacao=f.idlocalizacao AND j.idconflito=f.idconflito
4 AND f.idconflito <> 1
5 GROUP By pais

```

	Data Output	Messages	Notifications	Explain
	classificacao double precision	pais character varying (200)		
1	13.375	Turkey		
2	7.714285714285714	Italy		
3	8.541666666666666	Russia		
4	5	Sweden		
5	14.666666666666666	Norway		
6	7	Armenia		

Figura 84: Código SQL utilizado e respetivos resultados obtidos

Para se conseguir obter uma melhor percepção dos resultados obtidos, importou-se as tabelas resultante das *queries* para o Power BI e visualizou-se os resultados através da visualização Gráfico de linhas, visível na figura seguinte. No eixo dos Xx encontram-se os países que participam no concurso. No eixo dos Yy encontram-se as classificações médias dos países quando se encontravam em conflito e quando já não se encontravam em conflito.

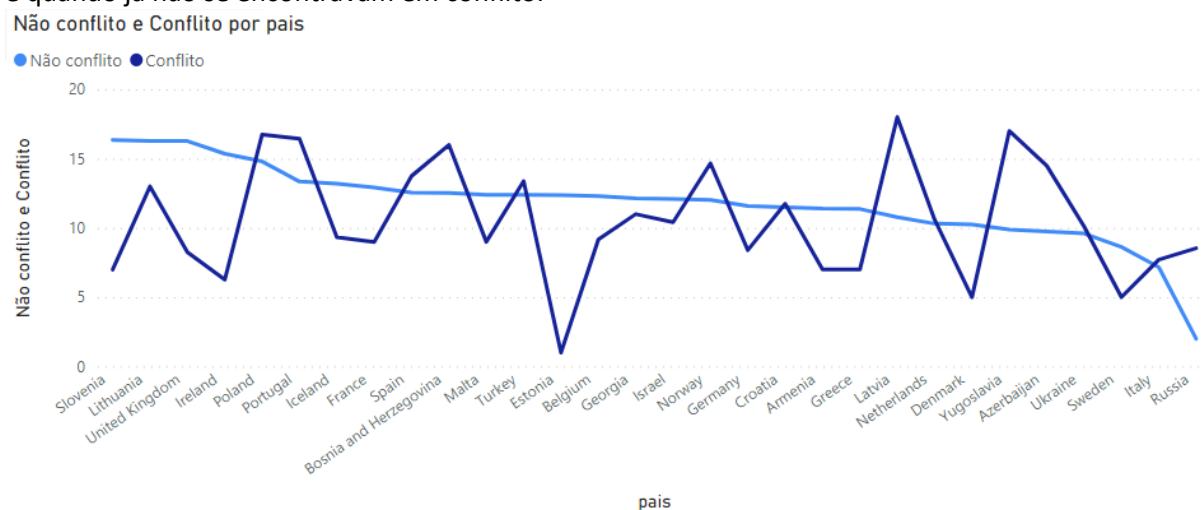


Figura 85: Comparação das médias das classificações obtidas por cada país tendo em conta se está perante um conflito ou não.

Ao observarmos a figura 85, é visível que há 13 países cuja sua classificação quando está perante um conflito é mais baixa – pior classificação - do que quando não está em conflito. No entanto, os restantes países apresentam melhores classificações no festival quando estão em conflito. Contudo, não conseguimos justificar qual dos países é que originou o conflito devido à pouca informação disponível nos datasets.

### 11.3.2. Os países “verdes” são mais populares?

Para responder a esta pergunta, recorreu-se ao programa *Power BI* e utilizando a visualização Gráfico de linhas e colunas empilhadas, visível na figura seguinte. No eixo do Xx, encontram-se os países que participam na eurovisão, já no eixo dos Yy, encontra-se as emissões de CO<sub>2</sub>. A reta que se encontra na figura corresponde ao valor média da classificação de cada país.

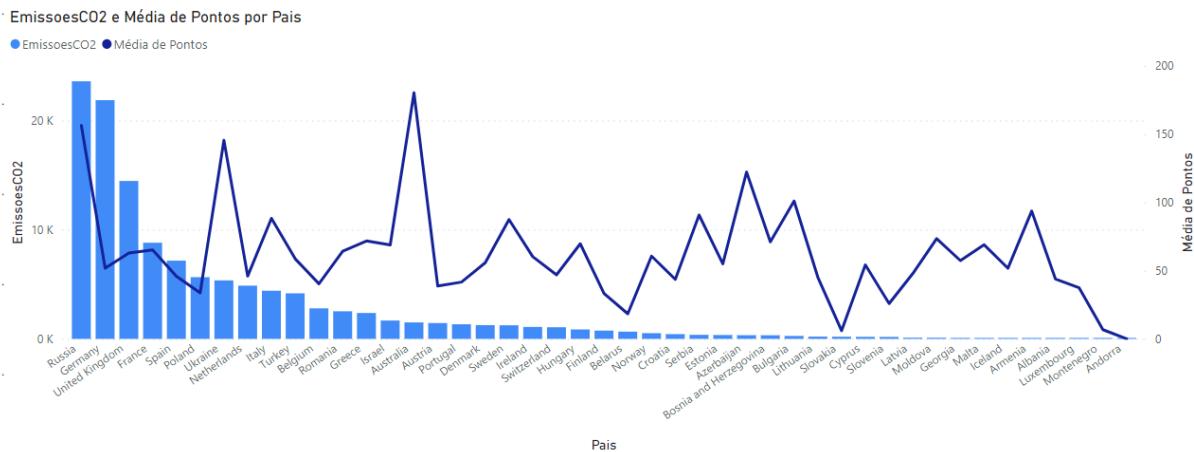


Figura 86: Comparação das médias das classificações obtidas por cada país com as emissões de CO<sub>2</sub> de cada país.

Ao observar a figura 86, conseguimos perceber que não há correlação das emissões de CO<sub>2</sub> com a classificação dos países no festival.

### 11.3.3 O turismo influencia a votação?

Para responder à questão utilizou-se, novamente, o *Power BI*, através da visualização Gráfico de linhas e colunas empilhadas, visível na figura seguinte. No eixo do Xx, encontram-se os países que participam na eurovisão, já no eixo dos Yy, encontra-se o número de turistas por área. A reta que se encontra na figura corresponde ao valor média da classificação de cada país.

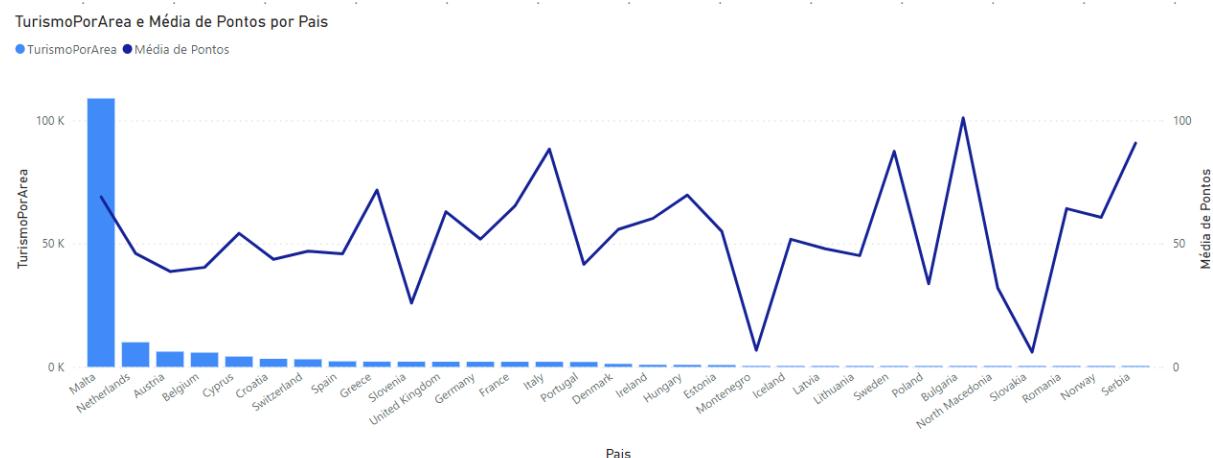


Figura 87: Comparação das médias das classificações obtidas por cada país com o turismo por área.

Ao observar a figura em cima, é possível verificar que não existe correlação entre o turismo e a classificação dos países no festival.

## 11.4 Curiosidades

Durante a prospeção de dados foram realizadas algumas interrogações sobre os dados que não faziam parte do âmbito das questões analíticas que tinham sido formuladas.

O resultado mais interessante, e que decidimos partilhar no relatório, é a influência da ordem de atuação no resultado. Este resultado não é relevante para o nosso processo de negócio, pois estas variáveis só são conhecidas próximo da data do concurso.

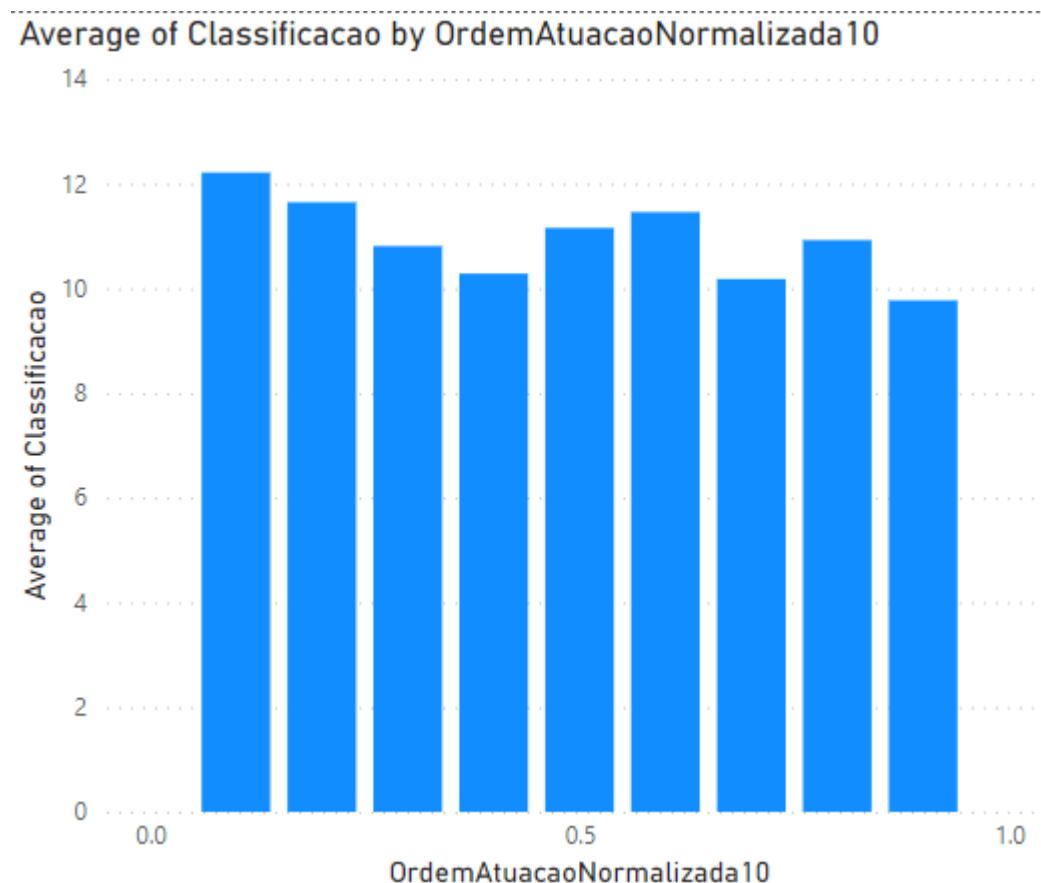


Figura 88 - Classificação média no concurso por ordem de atuação

Na figura acima podemos ver a variação da classificação média por ordem de atuação no concurso. Parece existir uma melhor classificação para músicas que tocam nos últimos 10% da parte final da Eurovisão.

Realizando uma análise para as décadas de 2000, 2010 e 2020, visíveis nas figuras representadas abaixo, a tendência mantém-se para os países que tocam no final, mas também imediatamente antes da metade do concurso.

**Classificacao by OrdemAtuacaoNormalizada10**

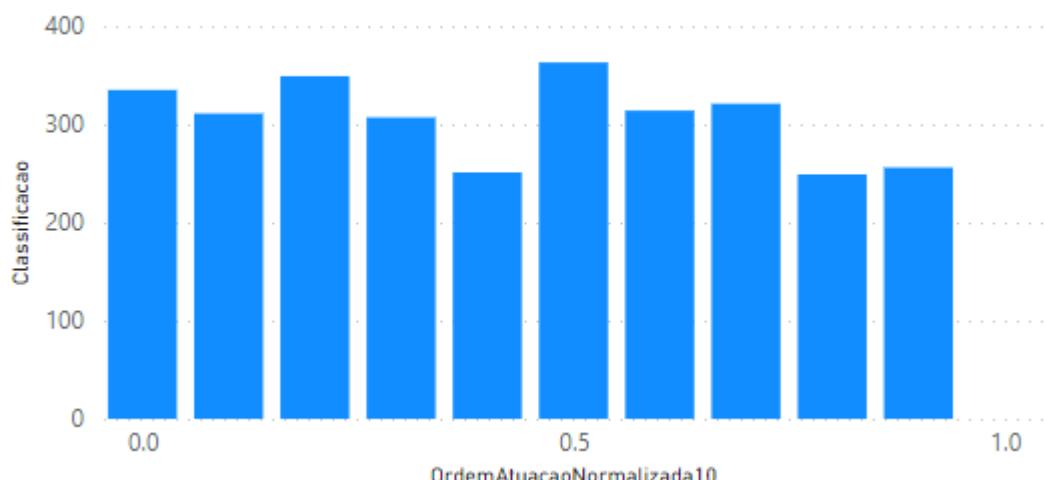


Figura 89 - Classificação média no concurso por ordem de atuação na década de 2000

**Classificacao by OrdemAtuacaoNormalizada10**

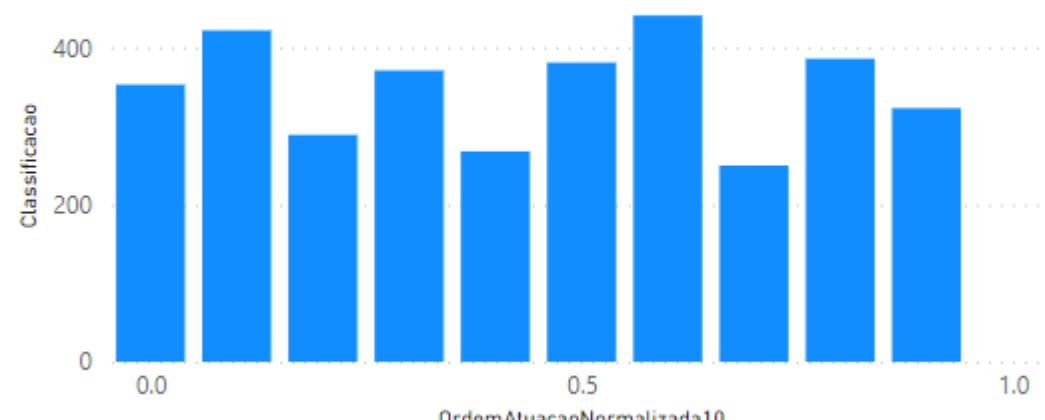


Figura 90 - Classificação média no concurso por ordem de atuação na década de 2010

**Classificacao by OrdemAtuacaoNormalizada10**

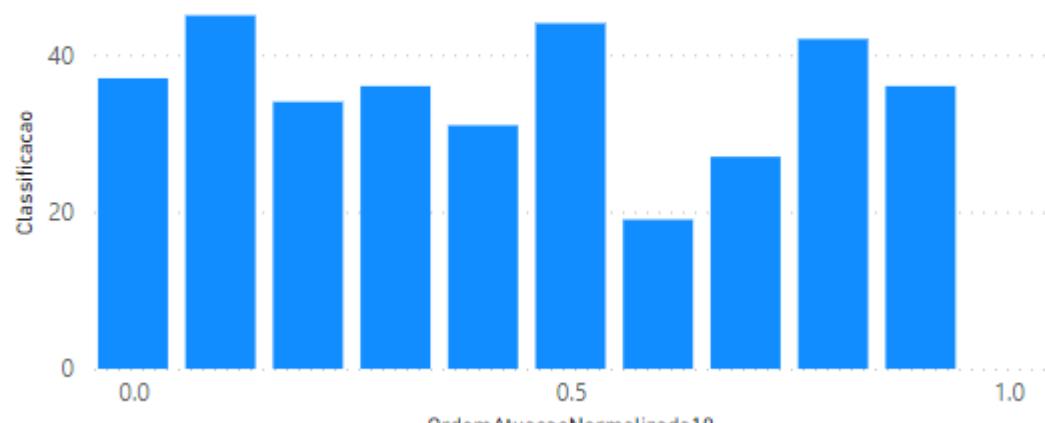


Figura 91 - Classificação média no concurso por ordem de atuação na década de 2020

## 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.

Na terceira etapa deste trabalho começámos por fazer pequenos ajustes ao trabalho que realizámos nas etapas anteriores, porém o foco principal manteve-se no desenvolvimento de um sistema ETL e na produção de relatórios para responder às perguntas analíticas identificadas na etapa 1. Para isso, implantámos automatismos para extrair dados dos repositórios, transformá-los na data *staging area* e carregar os dados tratados para a data *presentation area*.

Nesta etapa tivemos de abandonar algumas análises iniciais que pretendíamos fazer, como por exemplo o género de música, devido à falta de dados. Não obstante, seria muito interessante, em análises futuras, ver os géneros de música mais populares por país *versus* o televoto do país, assim como aprofundar a influência do turismo na Eurovisão, ou seja, ver o país de origem dos turistas *versus* o televoto, o número e origem de emigrantes *versus* televoto, etc.

Relativamente às perguntas analíticas e correspondentes respostas ficámos, por um lado satisfeitos, mas por outro não. Isto é, quanto à 1<sup>a</sup> pergunta analítica (Qual a influência da língua em que a canção é cantada?) conseguimos obter resultados convincentes sobre a existência de uma correlação entre a língua em que uma música é cantada e a pontuação recebida. Para a 2<sup>a</sup> pergunta analítica (Como é que a demografia e a geografia influenciam os resultados na eurovisão?) obtivemos resultados aceitáveis porque conseguimos perceber a influência da geografia (vizinhança) dos países participantes da Eurovisão nos resultados e identificar relações de entreajuda, inclusive identificamos que não existe nenhuma correlação entre a demografia de um país e a pontuação recebida. Na 3<sup>a</sup> pergunta analítica (As questões da atualidade influenciam os resultados?) apercebemo-nos que os dados que tínhamos relativos aos conflitos não eram suficientes para tirar boas conclusões e que não existe correlação entre os restantes fatores da atualidade.

## 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](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](https://en.wikipedia.org/wiki/Category:Lists_of_wars_by_date)