



PROJETO DE INTEGRAÇÃO E PROCESSAMENTO ANALÍTICO DE INFORMAÇÃO

Etapa 1 – Análise de Dados

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

Conteúdo

Índice de Figuras	2
Índice de Tabelas	3
Introdução	4
1. Fontes de dados	5
2. Descrição e análise dos datasets	7
2.1 Eurovision finals voting results 1957-2021	7
2.1.1 Análise estatística	8
2.2 Top Spotify 2010-2019	9
2.2.1 Análise estatística	9
2.3 Top spotify 2020-2021	10
2.3.1 Análise estatística	10
2.4 Land Borders (Neighbours).....	10
2.4.1 Análise estatística	11
2.5 Eurovision song lyrics 1956-2021	12
2.5.1 Erros e dados em falta.....	13
2.5.2 Análise estatística	14
2.6 Countries and Territories	19
2.6.1. Análise estatística.....	19
2.7 Tourists.....	20
2.7.1 Análise estatística	20
2.8 Historical Emissions	21
2.8.1 Análise estatística	22
2.9 World GDP (PIB 1960-2020)	23
2.9.1 Erros e dados em falta.....	24
2.9.2 Análise estatística	24
2.10 World Population (1960-2020)	25
2.10.1 Análise estatística.....	25
2.11 Conflicts Participants	26
2.11.1 Erros nos dados.....	27
2.11.2 Análise Estatística.....	27
3. Diagrama Relacional das Fontes de Fados	30
4. Processo de Negócio.....	31
5. Questões analíticas.....	32
Conclusão.....	33
Bibliografia	33

Índice de Figuras

Figura 1 - Moda de votos recebidos	8
Figura 2 - Moda de cidade anfitriã	8
Figura 3 - Moda de votos dados	8
Figura 4 - Contagem do número de vizinhos	11
Figura 5 - Países com mais participações na Eurovisão a partir do dataset música da Eurovisão	14
Figura 6 - Distribuição do género musical do dataset música da Eurovisão	15
Figura 7 - Músicas com pontuações mais altas no dataset música da Eurovisão	16
Figura 8 - Músicas com pontuações mais baixas no dataset música da Eurovisão	16
Figura 9 - Campo de validação da linguagem do dataset música da Eurovisão	16
Figura 10 - Linguagens mais utilizadas no dataset música da Eurovisão	17
Figura 11 - Países anfitriões do maior número de edições no dataset música da Eurovisão	18
Figura 12 - Cidades anfitriãs do maior número de edições no dataset música da Eurovisão	18
Figura 13 - Contagem dos países por continente	19
Figura 14 - Análise do campo "Number" por ano	20
Figura 15 - Extrato do gráfico do campo "Number" por país	21
Figura 16 - Análise do campo "Number" por país	23
Figura 17 - Análise do campo "Number" por ano	22
Figura 18 - PIB da Europa (em €) de 1970 a 2020	24
Figura 19 - Taxa de crescimento do PIB na Europa de 1970 a 2020	25
Figura 20 - População total em Portugal (1960-2020)	26
Figura 21 - População total na União Europeia (1960-2020)	26
Figura 22 - Contagem dos conflitos por State	28
Figura 23 - Países com mais participações em conflitos no dataset Conflicts Participants	29
Figura 24 - Localizações com maior ocorrência de conflitos no dataset Conflicts Participants	29
Figura 25 - Diagrama relacional entre tabelas	30
Figura 26 - Probabilidade de vencer a Eurovisão em 9/03/2022 e 31/03/2022 (EurovisionWorld, 2022)	32

Índice de Tabelas

Tabela 1 - Conjunto de dados e a sua descrição	5
Tabela 2 - Descrição dos campos do dataset da Eurovisão	7
Tabela 3 - Máximos e mínimos do dataset da Eurovisão.....	8
Tabela 4 - Modas do dataset da Eurovisão	8
Tabela 5 - Descrição dos campos do dataset Spotify 2010-2019.....	9
Tabela 6 - Moda e número de ocorrências para o género musical mais ouvido entre os anos 2010 e 2019.....	9
Tabela 7 - Descrição dos campos do dataset Spotify 2020-2021.....	10
Tabela 8 - Moda e número de ocorrências para o género musical mais ouvido em 2021	10
Tabela 9 - Descrição dos campos do dataset Land Borders	11
Tabela 10 - Análise estatística do campo No. Of Neighbours	11
Tabela 11 - Descrição dos campos do dataset música da Eurovisão.....	12
Tabela 12 - Descrição dos campos adicionados ao dataset música da Eurovisão	13
Tabela 13 - Moda do país do dataset música da Eurovisão	14
Tabela 14 - Moda do artista do dataset música da Eurovisão.....	14
Tabela 15 - Moda do género musical do dataset música da Eurovisão	15
Tabela 16 - Análise estatística das pontuações do dataset música da Eurovisão	15
Tabela 17 - Modas das linguagens do dataset música da Eurovisão	16
Tabela 18 - Análise estatística dos campos "value.pl" e "running order" do dataset música da Eurovisão	17
Tabela 19 - Análise estatística dos campos "Eurovision number" e "Year" do dataset música da Eurovisão	17
Tabela 20 - Modas das localizações do dataset música da Eurovisão	18
Tabela 21 - Descrição dos campos do dataset Countries and Territories	19
Tabela 22 - Descrição dos campos do dataset Tourists.....	20
Tabela 23 - Análise do campo Time.....	20
Tabela 24 - Descrição dos campos do dataset Historical Emissions.....	21
Tabela 25 - Análise do campo Year.....	22
Tabela 26 - Descrição dos campos do dataset World GDP	23
Tabela 27 - Análise estatística dos valores para a região da Europa	24
Tabela 28 - Descrição dos campos do dataset World Population.....	25
Tabela 29 - Análise estatística da população de Portugal e União Europeia (1960 e 2020)	25
Tabela 30 - Descrição dos campos do dataset Conflicts Participants	26
Tabela 31 - Descrição dos campos da tabela auxiliar do dataset Conflicts Participants.....	27
Tabela 32 - Descrição dos novos campos da tabela do dataset Conflicts Participants.....	27
Tabela 33 - Análise da Moda para o Campo "State"	27
Tabela 34 - Análise estatística do campo "Conflicts"	28
Tabela 35 - Descrição da extensão dos dados da tabela do dataset Conflicts Participants.....	28
Tabela 36 - Descrição da extensão dos dados da nova tabela do dataset Conflicts Participants	28
Tabela 37 - Descrição das modas dos dados da nova tabela do dataset Conflicts Participants.....	29

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 66ª 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.

1. Fontes de dados

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

Tabela 1- Conjunto de dados e a sua descrição

Dataset	Descrição	Formato	Extração	Link
Eurovision finals voting results 1957-2021	Dataset retirado do Kaggle com os votos de todos os países na Eurovisão de 1957 a 2021	.csv	Download direto	https://www.kaggle.com/orianao/eurovision-finals-voting-results-19572021
Top spotify (2010-2019)	Dataset retirado do Kaggle com as músicas mais ouvidas por ano (2010-2019) no mundo pelo Spotify, com dados baseados na Billboard	.csv	Download direto	https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year
Top spotify (2020-2021)	Dataset retirado do Kaggle com as músicas mais ouvidas do Spotify nos anos 2020 e 2021.	.csv	Download direto	https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021
Land Borders (Neighbours)	Lista de países e territórios por fronteiras terrestres. Inclui o número de fronteiras terrestres distintas de cada país ou território, bem como os nomes de seus países e territórios vizinhos.	.xlsx	Tabela de página convertida em tabela Excel	https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders
Eurovision song lyrics 1956-2021	Dataset retirado do Kaggle com dados sobre as músicas que participaram nos festivais da Eurovisão entre 1956 e 2021	.json	Download direto, importado e convertido em tabela Excel	https://www.kaggle.com/minitree/eurovision-song-lyrics

Countries and Territories	Dataset que exhibe os limites administrativos mundiais de nível 0. Contém uma lista dos países, bem como territórios não soberanos	.xlsx	Download direto	https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/table/
Tourist	Dataset retirado do Eurostat que exhibe resultados do número de chegadas de turistas a estabelecimentos de alojamento turístico por país (com base em hotéis e acomodações similares)	.xls	Download direto	https://ec.europa.eu/eurostat/databrowser/view/tour_occ_ar_nat/default/table?lang=en
Historical emissions	Dataset retirado do Kaggle com emissões históricas de dióxido de carbono ao longo de 3 décadas, por todos os países do mundo	.csv	Download direto	https://www.kaggle.com/datasets/ankan_hore545/carbon-dioxide-emissions-of-the-world
World GDP (PIB 1960-2020)	Dataset retirado do Kaggle com os valores do PIB, o crescimento do PIB, o PIB per capita e o crescimento do PIB per capita para todos os países do mundo entre 1960 e 2020.	.csv	Download direto	https://www.kaggle.com/datasets/zgrcemta/world-gdp-gdp-gdp-per-capita-and-annual-growths
World Population (1960-2020)	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	https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=WLD#

Conflicts Participants	Dataset retirado do Kaggle com uma lista de conflitos mundiais ocorridos depois do final da Segunda Guerra Mundial e os países envolvidos nesses conflitos	.csv	Download direto	https://www.kaggle.com/datasets/guybarash/war-conflicts-and-nations-who-took-part-in-them?select=Conflicts+participants.csv
-------------------------------	--	------	-----------------	---

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 novos campos para permitir uma melhor integração entre tabelas no futuro.

2.1 Eurovision finals voting results 1957-2021

O ficheiro “*EurovisionFinalsVotingResults1957-2021.csv*” 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
Index	Número inteiro	Identificador único	544
Host City	Texto	País anfitrião da Eurovisão	Tel aviv
Year	Número Inteiro	Ano em que se realizou a Eurovisão	2019
Points type	Texto	Tipo de pontos dados	Points given by televoters
From	Texto	O país que deu os pontos	Albania
To	Texto	O país que recebeu os pontos	Italy
Points	Número Inteiro	O 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 “ID”, “Year” e “Points”. O campo “ID” tem o máximo de 13446 que corresponde ao número total de instâncias de um país a dar pontos a outro país. O campo “Year” mostra a variação temporal dos dados, que vão desde 1957 até 2021. Finalmente o campo “Points” mostra o intervalo dos pontos possíveis, entre 1 e 12.

Tabela 3 - Máximos e mínimos do dataset da Eurovisão

Campo	Máximo	Mínimo
ID	13446	0
Year	2021	1957
Points	12	1

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. Através desta análise foi possível concluir que a cidade que mais vezes acolheu o festival foi Dublin, um total de 6 vezes. Também desta tabela é possível verificar que o país que mais vezes recebeu pontos foi a Suécia e 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.

Tabela 4 - Modas do dataset da Eurovisão

Campo	Moda	Ocorrências
Host City	Dublin	6
To	Sweden	721
From	United Kingdom	624

Count of To by To

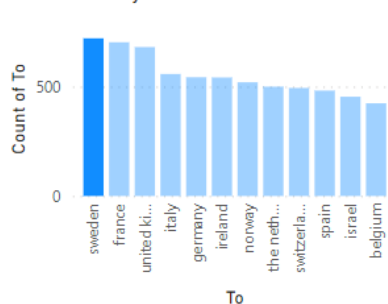


Figura 1 - Moda de votos recebidos

Count of Year by Host City

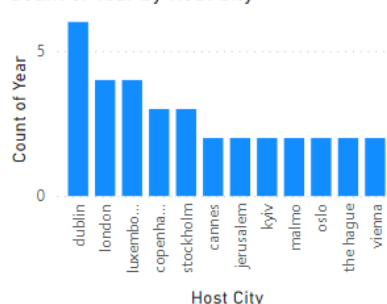


Figura 2 - Moda de cidade anfitriã

Count of From by From

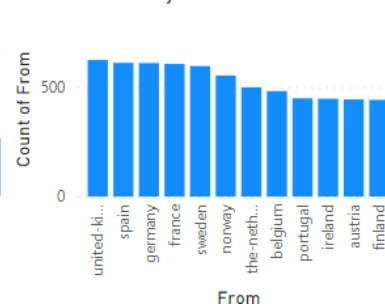


Figura 3 - Moda de votos dados

2.2 Top Spotify 2010-2019

Após realizar download do ficheiro “*TopSpotify2010-2019.csv*”, 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, eliminamos os campos que considerámos irrelevantes, ficando com os seguintes, estando estes representados na tabela abaixo.

Tabela 5 - Descrição dos campos do dataset Spotify 2010-2019

Campo	Tipo de dados	Descrição	Exemplo
Index	Número inteiro	Identificador único	1
Title	Texto	Nome da música	“Hey, Soul Sister”
Artist	Texto	Nome do artista da música	Train
Top genre	Texto	O género da música	Dance Pop
Year	Número inteiro	Ano da música no Billboard	2017
pop	Número inteiro	Popularidade	83

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.

2.2.1 Análise estatística

Para que cada ano tivesse o mesmo número de músicas mais ouvidas, reduzimos as amostras a 31 e realizámos a moda para cada ano, com o objetivo de perceber qual o género musical mais ouvido em cada ano.

Tabela 6 - Moda e número de ocorrências para o género musical mais ouvido entre os anos 2010 e 2019.

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

Ao observarmos 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.

2.3 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 TopSpotify2010-2019.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.

Tabela 7 - Descrição dos campos do dataset Spotify 2020-2021

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

2.3.1 Análise estatística

À semelhança dos dados anteriores, reduzimos as amostras a 31 e realizámos a moda para o ano 2021.

Tabela 8 - Moda e número de ocorrências para o género musical mais ouvido em 2021

Ano	2021
Moda	Pop e Hip Hop
Ocorrência	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.

2.4 Land Borders (Neighbours)

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 9 - Descrição dos campos do dataset Land Borders

Campo	Tipo de dados	Descrição	Exemplo
Country	Texto	Nome do País	Portugal
No. Of Neighbours	Número Inteiro	Número de vizinhos	1
Neighbours	Texto	Lista dos países vizinhos	Spain

2.4.1 Análise estatística

Ao efetuarmos uma análise estatística sobre o campo “No. Of Neighbours” usando uma função do Excel obtemos a seguinte tabela.

Tabela 10 - Análise estatística do campo No. Of Neighbours

Campo	Média	Mediana	Moda	Mínimo	Máximo	Contagem
Number of Neighbours	3.288557	3	0	0	16	201

Desta análise podemos detetar que o número total de países considerados são de 201, com uma média de 3 vizinhos por país, e verificar o máximo e mínimo de vizinhos que um país pode ter.

Para se ter uma ideia da quantidade de países que partilham o mesmo número de países vizinhos fizemos uma contagem do campo “No. Of Neighbours”, onde observámos os resultados obtidos.

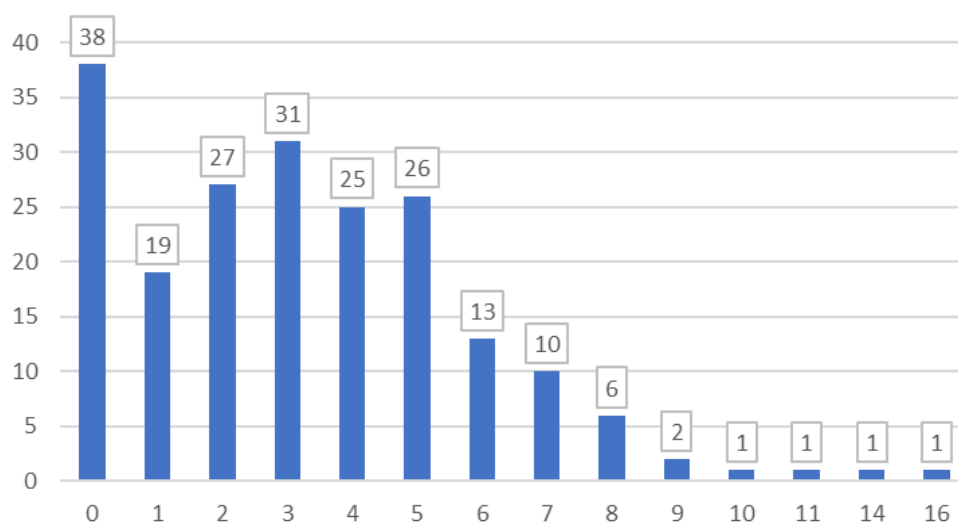


Figura 4 - Contagem do número de vizinhos

2.5 Eurovision song lyrics 1956-2021

O ficheiro “*Eurovision song lyrics 1956-2021.json*” 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.

Tabela 11 - Descrição dos campos do dataset música da Eurovisão

Campo	Tipo de dados	Descrição	Exemplo
Value. #	Número inteiro	Identificador único da música	176
Value.Country	Texto	Nome do País	Portugal
Value/#.1	Número inteiro	Número cumulativo de participações do país	5
Value.Artist	Texto	Nome do artista	Carlos Mendes
Value.Song	Texto	Nome da música	Verão
Value.Language	Texto	Linguagem da música	Portuguese
Value.Pl.	Número inteiro	Classificação Final	11
Value.Sc.	Número inteiro	Pontuação final	5
Value.Eurovision_Number	Número inteiro	Número de edição da eurovisão	13
Value.Year	Número inteiro	Ano da edição	1968
Value.Host_Country	Texto	País organizador	United Kingdom
Value.Host_City	Texto	Cidade onde foi realizado o festival	London
Value.Lyrics	Texto	Letra original da música	“Como tudo o que acaba Como pedra rolando duma fraga Como fumo subindo no ar” (...)
Value.Lyrics translation	Texto	Letra traduzida da música	“Like an ending to everything Like a stone rolling from a crag Like a smoke rising in the air” (...)

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 abaixo.

Tabela 12 - Descrição dos campos adicionados ao dataset música da Eurovisão

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

2.5.1 Erros e dados em falta

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

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

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

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

O campo “*ROpercent*” foi criado para normalizar a ordem das músicas com vista a facilitar a posterior análise dos dados. O problema está no facto de diferentes edições da Eurovisão terem sido realizadas com diferentes números de participantes, pelo que se pretendermos fazer uma análise como por exemplo “será que as músicas que tocam na segunda metade do concurso tem 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.

Existia ainda um erro na informação da tabela, o facto de esloveno aparecer como “slovene” e “slovenian”. Para o resolver, substituímos os valores de “slovenian” para “slovene”.

2.5.2 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.

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 13 - Moda do país do dataset música da Eurovisão

Campo	Moda	Value/#.1
Value.Country	Germany	64

Count of Value.Country by Value.Country

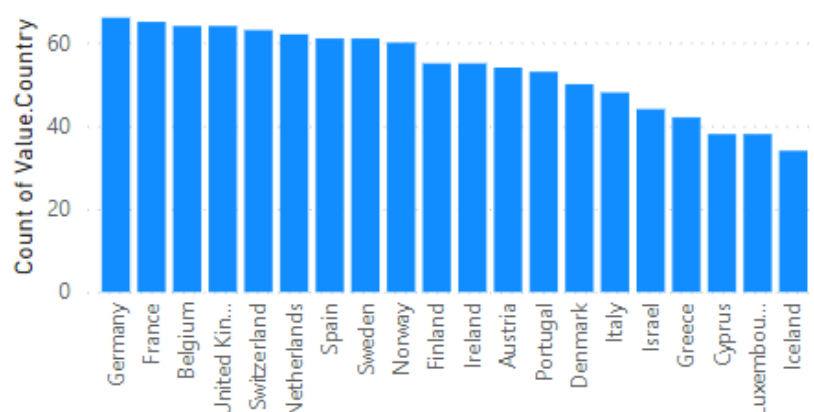


Figura 5 - Países com mais participações na Eurovisão a partir do dataset música da Eurovisão

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 14 - Moda do artista do dataset música da Eurovisão

Campo	Moda	Ocorrências	País Representado
Value.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 exceccionalmente cada país apresentou duas músicas.

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 15 - Moda do género musical do dataset música da Eurovisão

2010-2021	Moda	# Occurences
Genre	Pop	122

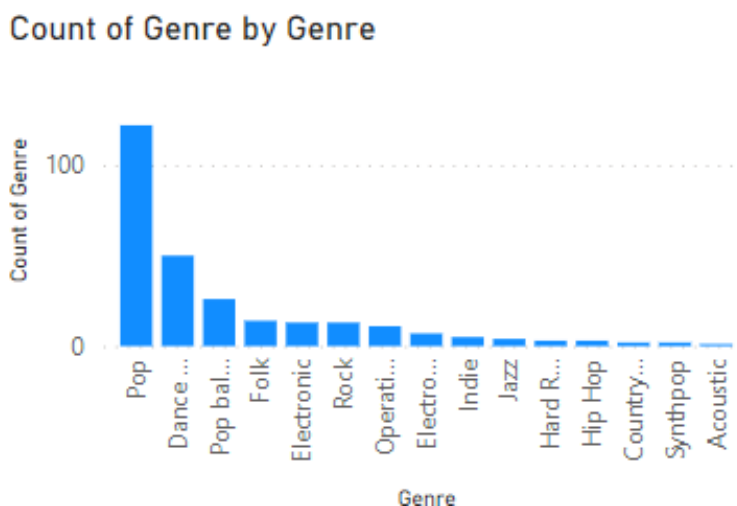


Figura 6 - Distribuição do género musical do dataset música da Eurovisão

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 16 - Análise estatística das pontuações do dataset música da Eurovisão

Campo	Média	Máximo	Mínimo	Desvio Padrão
Value.Sc	71	758	0	81

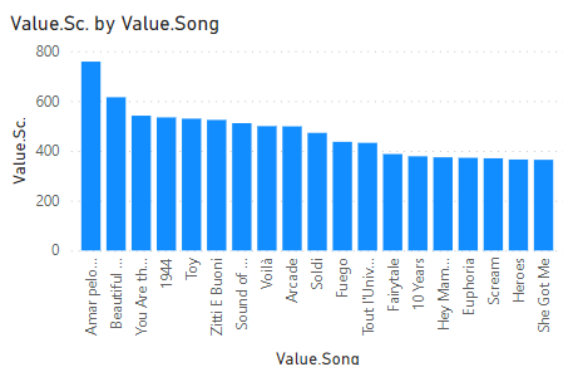


Figura 7 - Músicas com pontuações mais altas no dataset música da Eurovisão

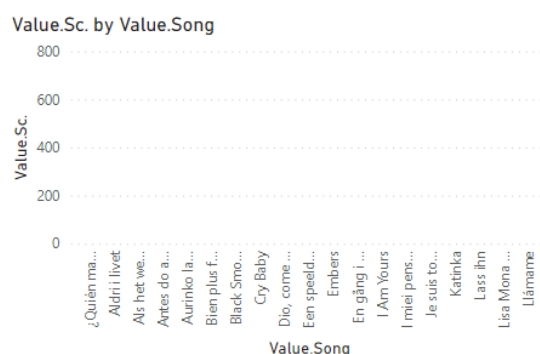


Figura 8 - Músicas com pontuações mais baixas no dataset música da Eurovisão

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, no entanto, existe um maior número de ocorrências somadas de outras linguagens.

Tabela 17 - Modas das linguagens do dataset música da Eurovisão

Campo	Moda	Ocorrências
Value.Language	English	615
EnglishNonEnglish	Not English	844

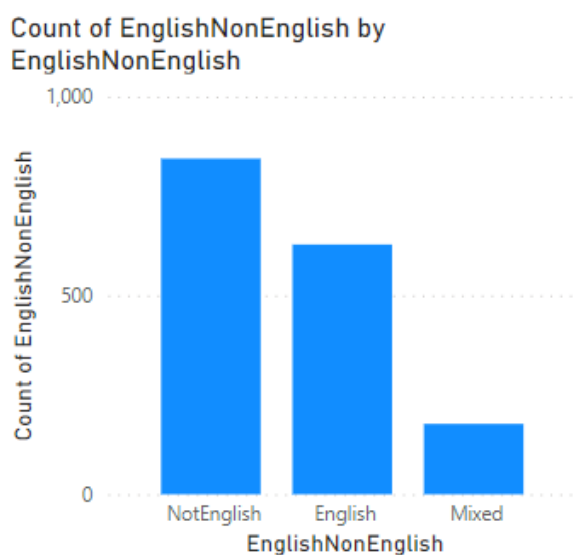


Figura 9 - Campo de validação da linguagem do dataset música da Eurovisão

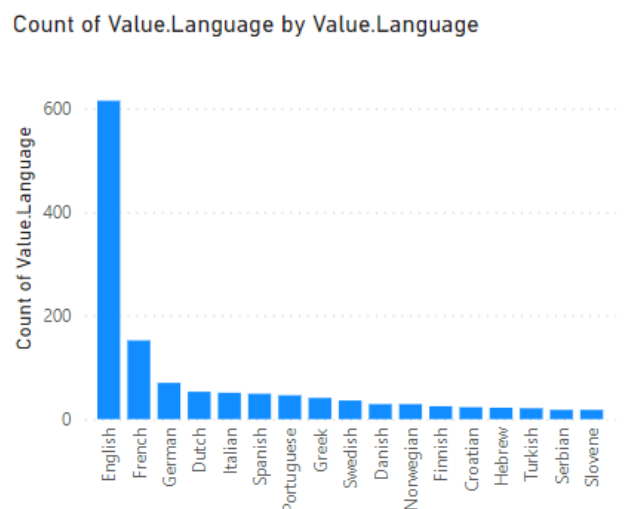


Figura 10 - Linguagens mais utilizadas no dataset música da Eurovisão

De seguida foram também analisados os campos de “value.pl” e “running order” em conjunto, de forma a ser mais fácil 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 18 - Análise estatística dos campos “value.pl” e “running order” do dataset música da Eurovisão

Campo	Media	Máximo	Mínimo	Desvio Padrão
Value.PI	11.26	26	1	6.68
Running Order	11.34	27	1	6.67

No que toca aos campos “Eurovision number” e “Year”, as funções, máximo e mínimo permitem perceber a extensão dos dados incluídos nesta dataset, que englobam as 65 edições do festival entre 1956 e 2021.

Tabela 19 - Análise estatística dos campos “Eurovision number” e “Year” do dataset música da Eurovisão

Campo	Máximo	Mínimo
Eurovision number	65	1
Year	2021	1956

Finalmente, a última análise 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 20 - Modas das localizações do dataset música da Eurovisão

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

Count of Value.Year by Value.Host_Country

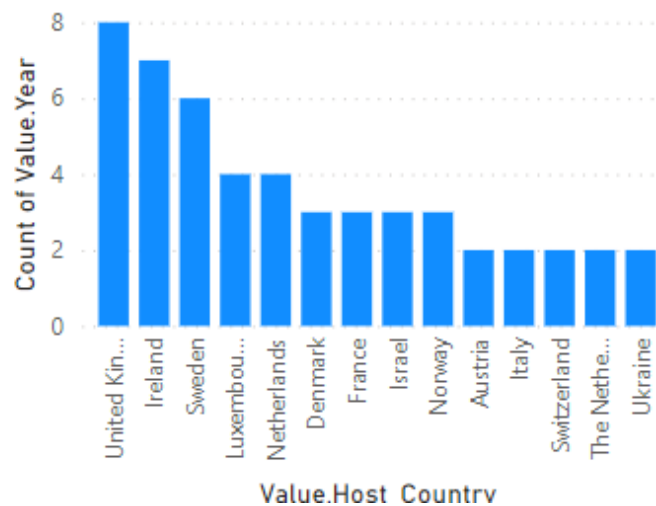


Figura 11 - Países anfitriões do maior número de edições no dataset música da Eurovisão

Count of Value.Year by Value.Host_City

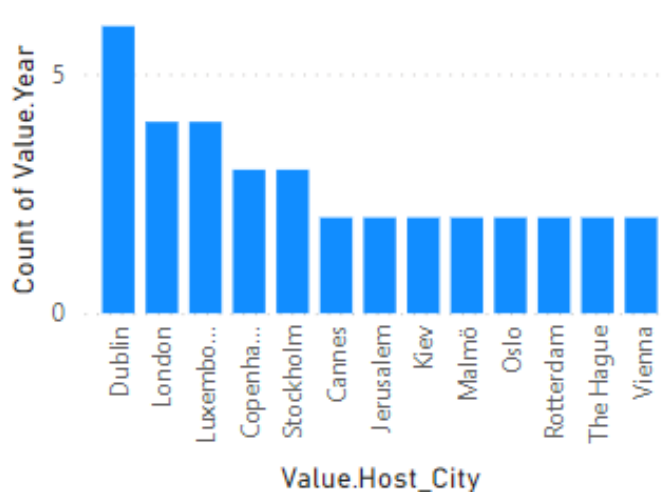


Figura 12 - Cidades anfitriãs do maior número de edições no dataset música da Eurovisão

2.6 Countries and Territories

A tabela “world-administrative-boundaries.xls”, tal como importada, incluía alguns campos que não eram pertinentes para a análise em estudo por isso procedemos à eliminação das mesmas.

Em baixo apresentamos os dados e informações mais relevantes da tabela world-administrative-boundaries.xls.

Tabela 21 - Descrição dos campos do dataset Countries and Territories

Campo	Tipo de dados	Descrição	Exemplo
ID	Número Inteiro	Identificador único do país	37
Country	Texto	Nome do País	Portugal
Alpha-2_Code	Texto	Código de duas-letras definidas no ISO 3166-1 para representar países.	PT
Continent of the territory	Texto	Designação continente a que pertence o país.	Europe
Region of the territory	Texto	Designação da região do país dentro do continente.	Southern Europe

2.6.1. Análise estatística

De modo a analisar a tabela foi estudada a moda dos continentes, o que se traduz no número de países presentes por continente no dataset.

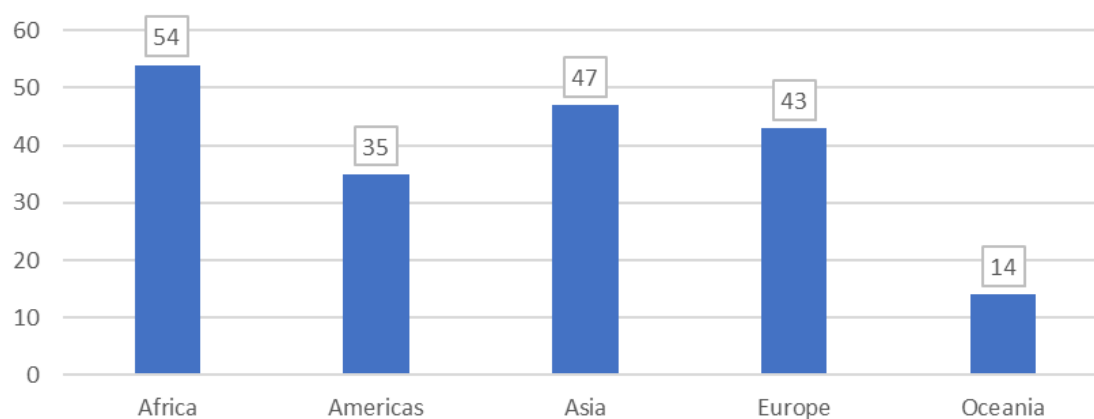


Figura 13 - Contagem dos países por continente

Ao realizar a análise estatística descritiva desta tabela concluímos que o número total de países é 193 e que para os 5 continentes considerados existe uma média de 38.6 países por continente. O continente com mais países é África (54) e o continente com menos países é a Oceânia (14).

2.7 Tourists

O ficheiro “*Tourists.xls*”, após o download, continha os seguintes campos, visíveis na tabela seguinte.

Tabela 22 - Descrição dos campos do dataset *Tourists*

Campo	Tipo de dados	Descrição	Exemplo
Geo (labels)	Texto	País de destino dos turistas	Belgium
Time	Número Inteiro	Ano a que se refere	2012
Number	Número Inteiro	Número de turistas	9041000

2.7.1 Análise estatística

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

Tabela 23 - Análise do campo *Time*

Campo	Mínimo	Máximo
Time	1990	2019

Analisando o campo “*Number*”, realizou-se um estudo estatístico, com o valor máximo, mínimo e a média por ano e por país, observável nos gráficos seguintes.



Figura 14 - Análise do campo “*Number*” por ano

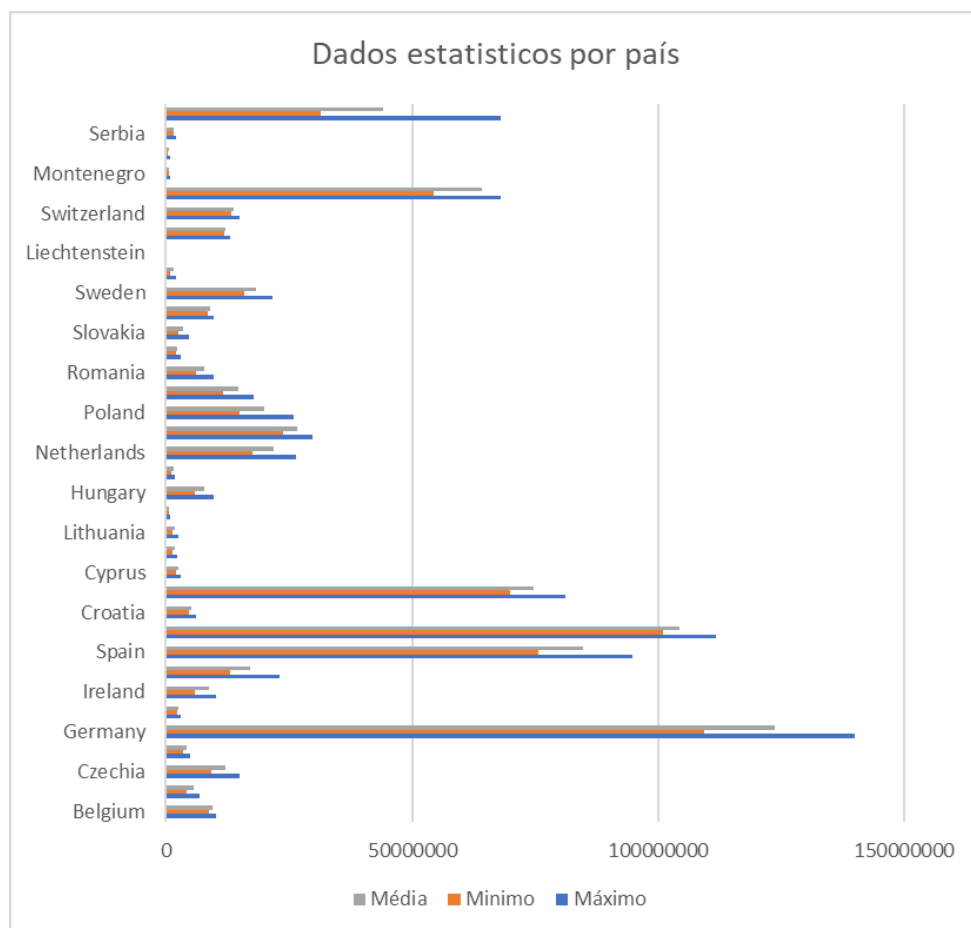


Figura 15 – Extrato do gráfico do campo "Number" por país

Ao observarmos a figura 14 conseguimos concluir que a atividade turista ao longo do tempo tem vindo aumentar gradualmente. Já na figura 15 é visível que o país que recebe mais turistas é a Alemanha e o que recebe menos é o Liechtenstein.

2.8 Historical Emissions

O ficheiro "*historical_emissions.csv*" guarda os dados de emissão de CO₂ de cada país no mundo. Este documento é constituído pelos seguintes campos.

Tabela 24 - Descrição dos campos do dataset Historical Emissions

Campo	Texto	Descrição	Exemplo
Country	Texto	Nome do país	Italy
Gas	Texto	Tipo de gás emitido	CO2
Unit	Texto	Unidade de medida	MtCO ₂ e
Year	Número inteiro	Ano a que se refere a emissão	2000
Dados anuais	Números Reais	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

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

Tabela 25 - Análise do campo Year

Campo	Mínimo	Máximo
Year	1990	2019

À semelhança do dataset anterior, realizámos, novamente, a análise estatística da média, do valor máximo e do valor mínimo de emissões por ano e por país. Reparámos, ainda, ao realizar a análise dos dados que existiam valores negativos. Em certos casos podem existir emissões de CO2 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.

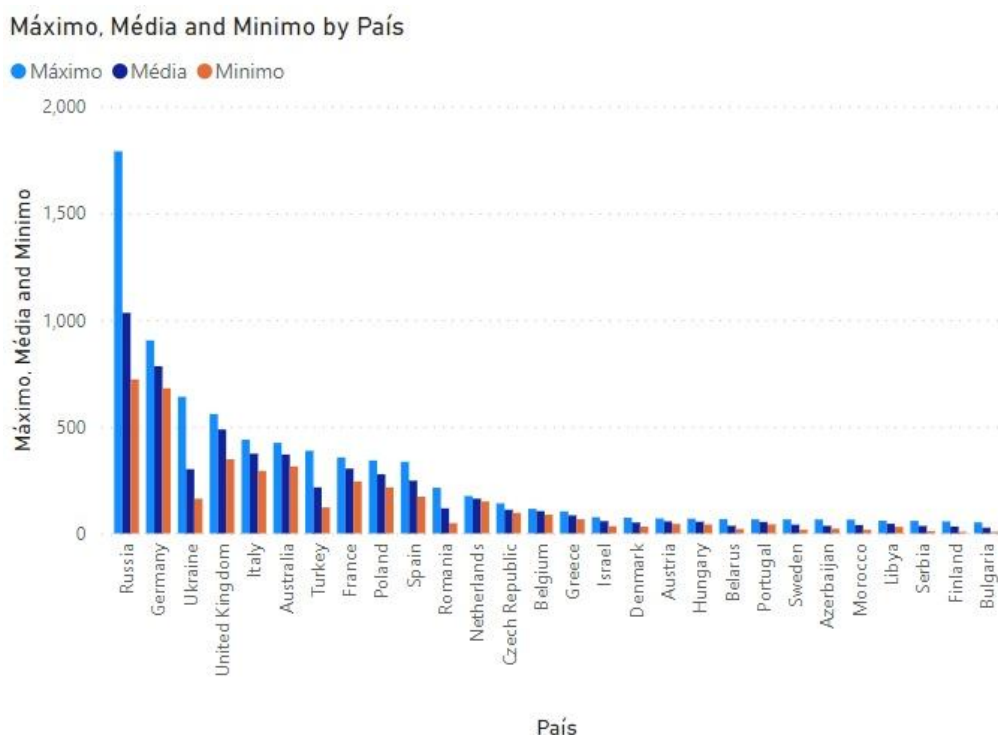


Figura 166 - Análise do campo “Number” por país

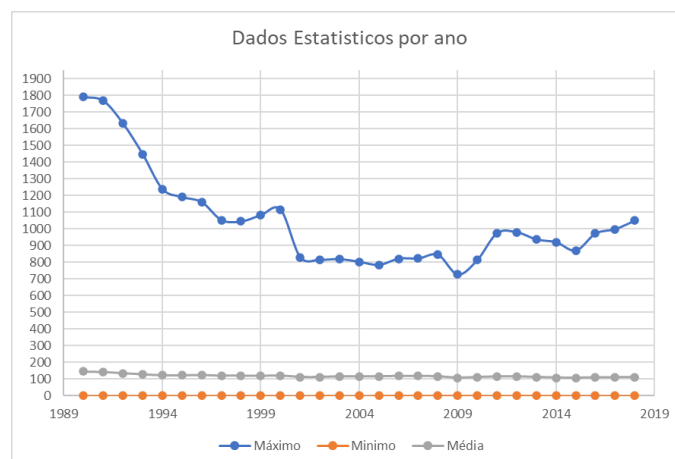


Figura 177 - Análise do campo "Number" por ano

Ao observar a figura “Dados estatísticos por ano” conseguimos determinar que o ano em que houve mais emissões de CO₂ 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 CO₂ 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 CO₂ é a Rússia e que o país que emite menos é Andorra.

2.9 World GDP (PIB 1960-2020)

O dataset extraído, referente ao Produto Interno Bruto (PIB), incluía 6 ficheiros no formato *.csv com uma estrutura igual à definida na tabela seguinte. Em cada um dos ficheiros temos dados anuais desde 1960 até 2020.

Tabela 26 - Descrição dos campos do dataset World GDP

Campo	Tipo de dados	Descrição	Exemplo
Country	Texto	Nome do País ou da Região	Portugal
Alpha-3_Code	Texto	Código de três-letras definidas no ISO 3166-1 para representar países.	PRT
Year	Número Real	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),	228,539,245,045.341 € -8.44246 % 22,176.3 € -8.61393% 351,924,219,905.62 € 34,148.95 €

2.9.1 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, iremos apenas considerar um intervalo de tempo em que tenhamos dados de todos os países.

2.9.2 Análise estatística

Para além das tabelas conterem uma lista de todos os países do mundo, estas incluem também dados para algumas regiões (agregação de países) por isso realizámos uma análise estatística de cada um dos valores para a região da Europa. É importante referir que para esta região só estão disponíveis dados a partir de:

- 1970 para os valores do PIB e PIB per capita,
- 1971 para o crescimento do PIB e do PIB per capita,
- 1990 para a PPC e PPC per capita.

Tabela 27 - Análise estatística dos valores para a região da Europa

Estatística	PIB	Crescimento PIB (%)	PIB per capita	Crescimento PIB per capita (%)	PIB PPC	PIB PPC per capita
Média	7,959,044,40 0,721.5 €	2.031929	18,351.8 €	1.730168	12,417,369,2 40,266.0 €	28,353.3 €
Mínimo	725,931,117, 202.7 €	-5.95563	1,879.1 €	-6.08266	6,227,973,80 2,592.0 €	14,811.7 €
Máximo	16,240,620,0 39,269.4 €	6.009084	36,920.8 €	5.327612	20,602,233,3 47,586.6 €	46,069.8 €
Contagem	51	50	51	50	31	31

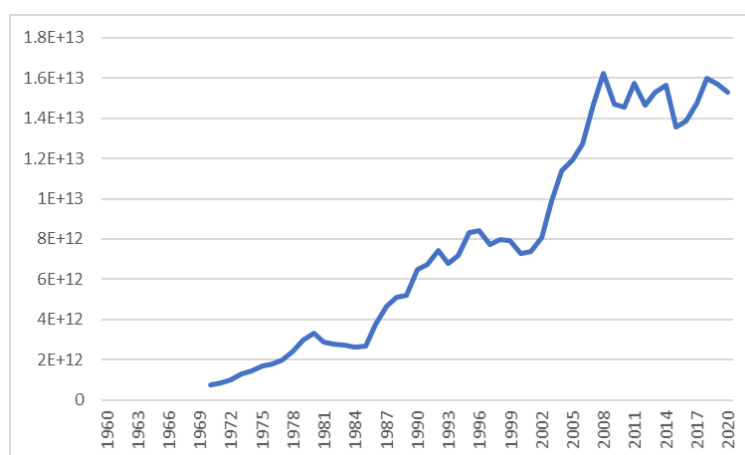


Figura 18 - PIB da Europa (em €) de 1970 a 2020

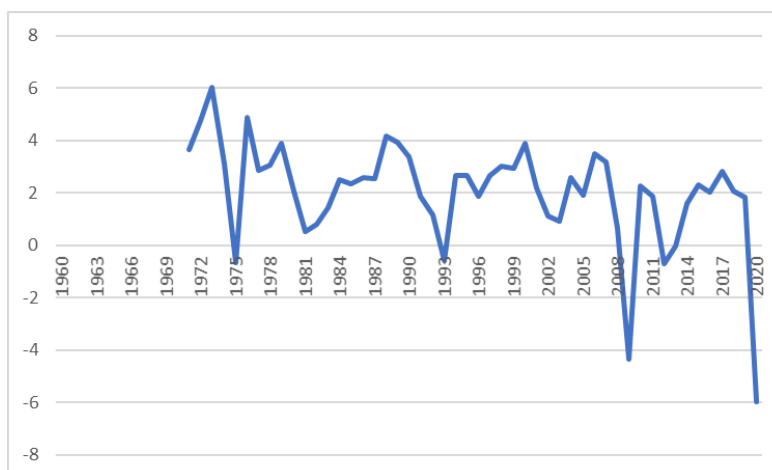


Figura 19 - Taxa de crescimento do PIB na Europa de 1970 a 2020

2.10 World Population (1960-2020)

A tabela “*WorldPopulation_Dataset.xlsx*” contém o número de população total anual para uma lista de países e regiões do mundo de 1960 a 2020.

Tabela 28 - Descrição dos campos do dataset *World Population*

Campo	Tipo de dados	Descrição	Exemplo
Country	Texto	Nome do País ou da Região	Portugal
Alpha-3_Code	Texto	Código de três-letras definidas no ISO 3166-1 para representar países.	PRT
Year	Número inteiro	Valor da população total	10,305,564

2.10.1 Análise estatística

No presente ponto realizámos uma análise estatística da população de Portugal e da População da União Europeia entre 1960 e 2020. Os resultados obtidos encontram-se na tabela seguinte.

Tabela 29 - Análise estatística da população de Portugal e União Europeia (1960 e 2020)

Estatística	Portugal	União Europeia
Média	9,816,538	414,770,477
Mínimo	8,630,430	356,906,098
Máximo	10,573,100	447,801,418

Nos gráficos seguintes observa-se a variação da população ao longo do tempo em Portugal e na União Europeia entre o ano de 1960 e o de 2020.

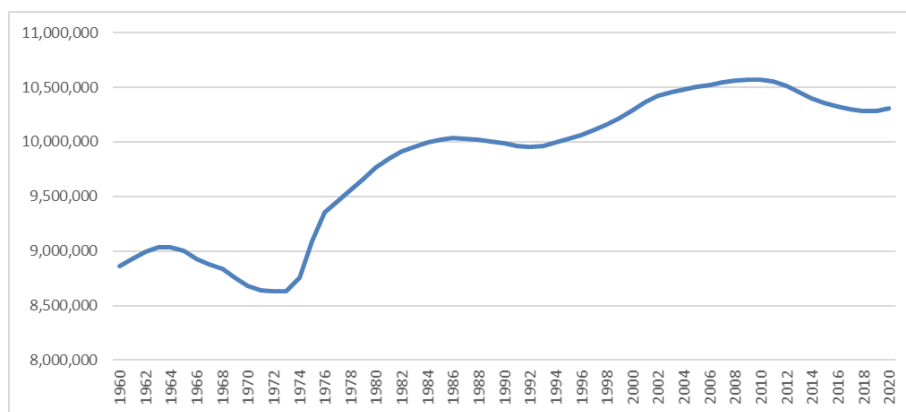


Figura 20 - População total em Portugal (1960-2020)

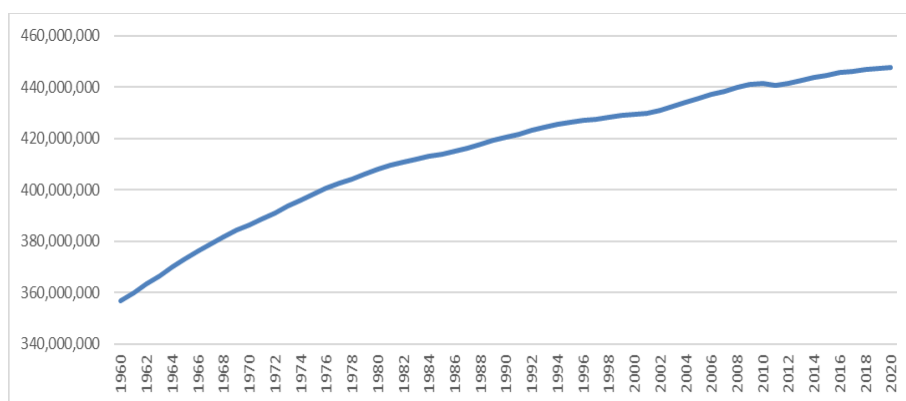


Figura 21 - População total na União Europeia (1960-2020)

2.11 Conflicts Participants

O ficheiro "*Conflictsparticipants.csv*" 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 30 - Descrição dos campos do dataset Conflicts Participants

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

Este dataset contém ainda uma tabela auxiliar que melhor permite realizar operações sobre os dados

Tabela 31 - Descrição dos campos da tabela auxiliar do dataset *Conflicts Participants*

Campo	Texto	Descrição	Exemplo
State	Texto	Nome do país	Pakistan
Conflicts	Inteiro	Número de conflitos em que esteve envolvido	9
Conflict1 (...) Conflict14	Texto	Nome do conflito	Second Congo War

2.11.1 Erros nos dados

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 nova tabela Excel que inclui os seguintes dados:

Tabela 32 - Descrição dos novos campos da tabela do dataset *Conflicts Participants*

Campo	Texto	Descrição	Exemplo
ID	Inteiro	Nome do país	1
Conflict Location	Texto	País onde o conflito se realizou maioritariamente	Morocco
Conflict	Texto	Nome do Conflito	Ifni War
Start Date	Data	Data de início	23/10/1957
End Date	Data	Data de fim	30/06/1958
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.

2.11.2 Análise Estatística

Utilizando a tabela original realizou-se uma análise estatística sobre os seus diversos campos. No que diz respeito ao campo “*State*”, observa-se na tabela seguinte que o país que participou em mais conflitos é os Estados Unido da América. De seguida procedeu-se a uma contagem, no Excel, dos conflitos por país, o que originou o gráfico seguinte.

Tabela 33 - Análise da Moda para o Campo “*State*”

Campo	Moda	Ocorrências
State	United States of America	14

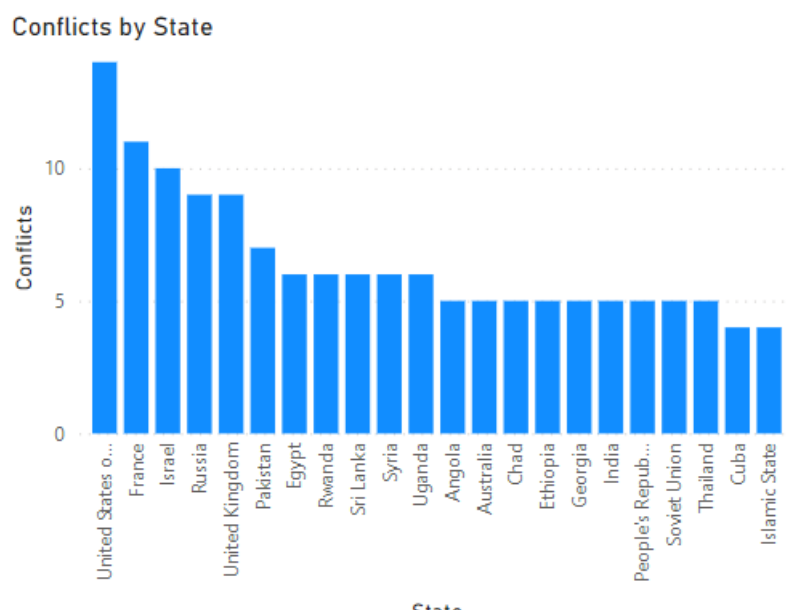


Figura 22 - Contagem dos conflitos por State

Olhando para o gráfico anterior decidimos realizar uma análise sobre o campo “Conflicts” onde analisamos o número máximo e mínimo de participações que cada país teve em conflitos, adicionalmente calculamos também a média e moda.

Tabela 34 - Análise estatística do campo “Conflicts”

Campo	Média	Máximo	Mínimo	Moda	Desvio Padrão
Conflicts	2.54	14	1	1	2.18

Como é visível na tabela seguinte, existem lacunas nos dados, estando as datas de início e as datas de fim repetidas, razão pela qual os dados tiveram de ser corrigidos e atualizados.

Tabela 35 - Descrição da extensão dos dados da tabela do dataset Conflicts Participants

Campo	Max	Min
Start date	2015-01-01T00:00:00Z	1946-01-01T00:00:00Z
End date	2015-01-01T00:00:00Z	1946-01-01T00:00:00Z

A análise estatística do dataset modificado tem alguns valores algo diferentes do dataset original. Existem agora 185 instâncias de conflito na tabela, variando as suas datas de início entre 1957 e 2022, sendo que existem conflitos que ainda não foram concluídos.

Tabela 36 - Descrição da extensão dos dados da nova tabela do dataset Conflicts Participants

Campo	Max	Min
ID	185	1
Start date	24/02/2022	23/10/1957
End date	21/05/2021	30/06/1958

Relativamente aos dados texto, foi possível procurar os valores mais comuns para localização do conflito, nomeadamente a Faixa de Gaza onde existem 7 ocorrências e o país que participou em mais conflitos durante este período de tempo, a França, com um total de 18.

Tabela 37 - Descrição das modas dos dados da nova tabela do dataset Conflicts Participants

Campo	Moda	# Ocorrências
Conflict Location	Gaza Strip	7
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.

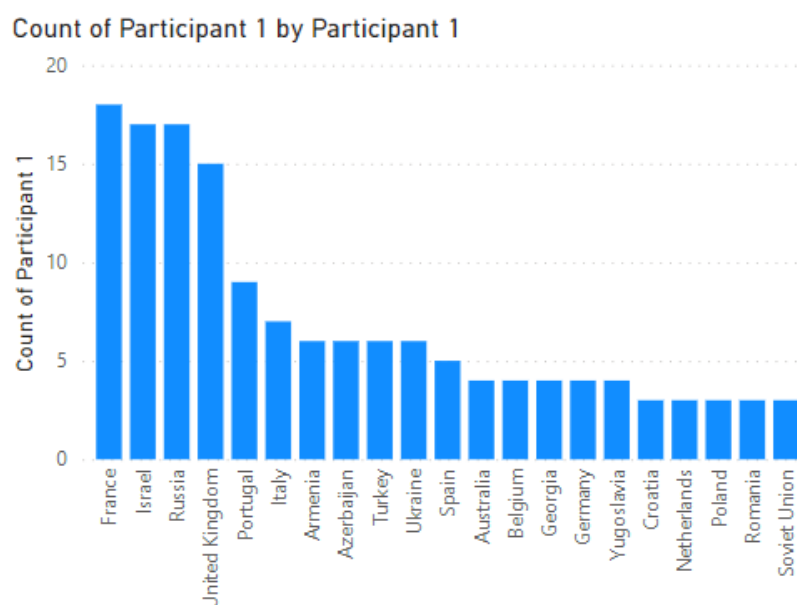


Figura 23 - Países com mais participações em conflitos no dataset Conflicts Participants

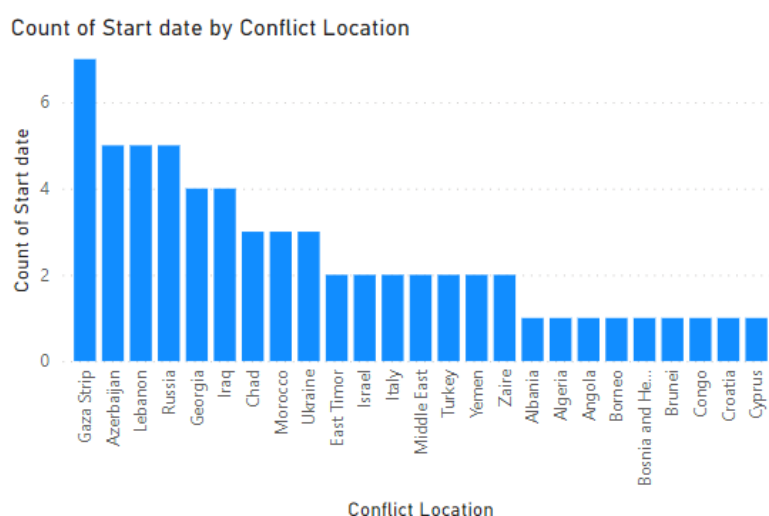


Figura 24 - Localizações com maior ocorrência de conflitos no dataset Conflicts Participants

3. Diagrama Relacional das Fontes de Fados

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

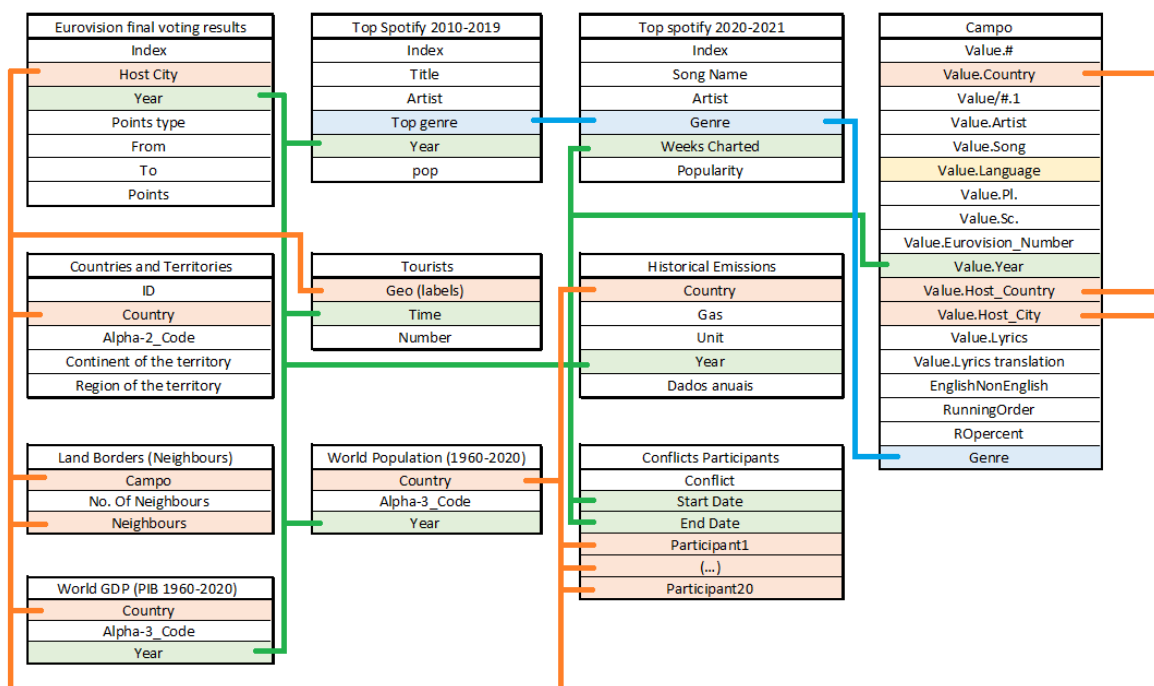


Figura 25 - Diagrama relacional entre tabelas

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.

Esta última cor não apresenta, nesta etapa, mais ligações com outros campos de outros conjuntos de dados, no entanto, é possível realizar, como por exemplo, com a tabela *Countries and Territories*, bastando, apenas, criar uma coluna com a língua oficial de cada país.

Tal acontece, pois, o grupo ainda não definiu se essa ligação será pertinente para o estudo sendo definido numa fase mais avançada do projeto.

4. Processo de Negócio

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

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

- Países
- Casas de apostas

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

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

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

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

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

		winning chance	BET365	SMARKETS*	BETSSON			winning chance	BET365	UNIBET	888 SPORT
1	Ukraine ♥️♥️ Kalush Orchestra - Stefa...	28%	2.37	2.94	2.3	1	Ukraine ♥️♥️ Kalush Orchestra - Stefa...	33%	6/5	6/5	6/7
2	Italy Mahmood & Blanco - Brividi	18%	3.5	3.8	3.95	2	Italy Mahmood & Blanco - Brividi	17%	10/3	13/4	3/1
3	Sweden	6%	12	15	11	3	Sweden Cornelia Jakobs - Hold Me Closer	12%	5/1	19/4	10/3
4	Greece artist: Amanda Tenfjord	5%	12	15	13	4	United Kingdom Sam Ryder - Space Man	5%	14/1	17/1	13/1
5	Poland Krystian Ochman - River	4%	15	24	18	5	Greece Amanda Tenfjord - Die Together	4%	18/1	20/1	14/1
6	United Kingdom	4%	17	20	17	6	Poland Ochman - River	4%	18/1	20/1	14/1
7	Norway Subwoolfer - Give That Wolf a Ba...	3%	23	48	20	7	Spain Chanel - SloMo	2%	25/1	33/1	28/1
8	Netherlands S10 - De Diepte	3%	26	32	25	8	Norway Subwoolfer - Give That Wolf a Ba...	2%	33/1	30/1	32/1
9	Belgium artist: Jérémie Makiese	2%	26	48	25	9	Netherlands S10 - De Diepte	2%	33/1	30/1	27/1
10	Spain Chanel - SloMo	2%	34	44	30	10	Australia Sheldon Riley - Not the Same	2%	40/1	50/1	39/1
11	Australia Sheldon Riley - Not the Same	2%	29	48	30	11	Portugal Maro - Saudade, saudade	1%	50/1	50/1	45/1
12	France Alvan & Ahez - Fulenn	2%	36	44	35	12	Belgium Jérémie Makiese - Miss You	1%	66/1	50/1	50/1
13	Switzerland Marius Bear - Boys Do Cry	2%	21	160	30	13	France Alvan & Ahez - Fulenn	1%	66/1	50/1	66/1
14	Finland The Rasmus - Jezebel	2%	51	100	40	14	Switzerland Marius Bear - Boys Do Cry	1%	66/1	66/1	66/1
15	Cyprus Andromache - Ela	2%	41	65	50	15	Serbia Konstrakta - In Corpore Sano	1%	66/1	60/1	74/1

Figura 26 - Probabilidade de vencer a Eurovisão em 9/03/2022 e em 31/03/2022 (EurovisionWorld, 2022)

5. Questões analíticas

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

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

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.

Bibliografia

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

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

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

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

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

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

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

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