

AD - TRABALHO PRÁTICO



Universidade do Minho
Escola de Engenharia

Grupo 4:

Carolina Resende Marques, PG42818

Francisco Borges, PG42829

Rui Pereira, PG42853

Vasco António Lopes Ramos, PG42852

Conteúdo

1.

Análise e Escolha da Fonte de Dados

2.


Arquitetura do Data Warehouse

3.

Processos de ETL

4.

Sistema de BI



01

Análise e Escolha da Fonte de Dados

Escolha da Fonte de Dados

- Na escolha do dataset houve a preocupação de garantir que o tamanho deste era suficientemente grande não só para garantir que a informação era relevante e útil, mas também que fosse possível construir uma análise adequada.
- A escolha recaiu sobre informações relacionadas com as músicas, artistas e géneros distribuídos na plataforma Spotify desde 1921 até 2020.

Ferramentas

Processo ETL



Desenvolvimento do DW



Business Intelligence





02

Arquitetura do Data Warehouse

Modelação

- Na construção do nosso data warehouse partimos com o objetivo de perceber quais os fatores que influenciam o mercado da música.
- Percebemos então que o nosso foco seriam as músicas, artistas e o géneros musicais.
- Achamos então que seria benéfico explorar ao máximo os dados fornecidos pelo dataset para que obtivéssemos uma maior flexibilidade analítica.

Dimensões selecionadas

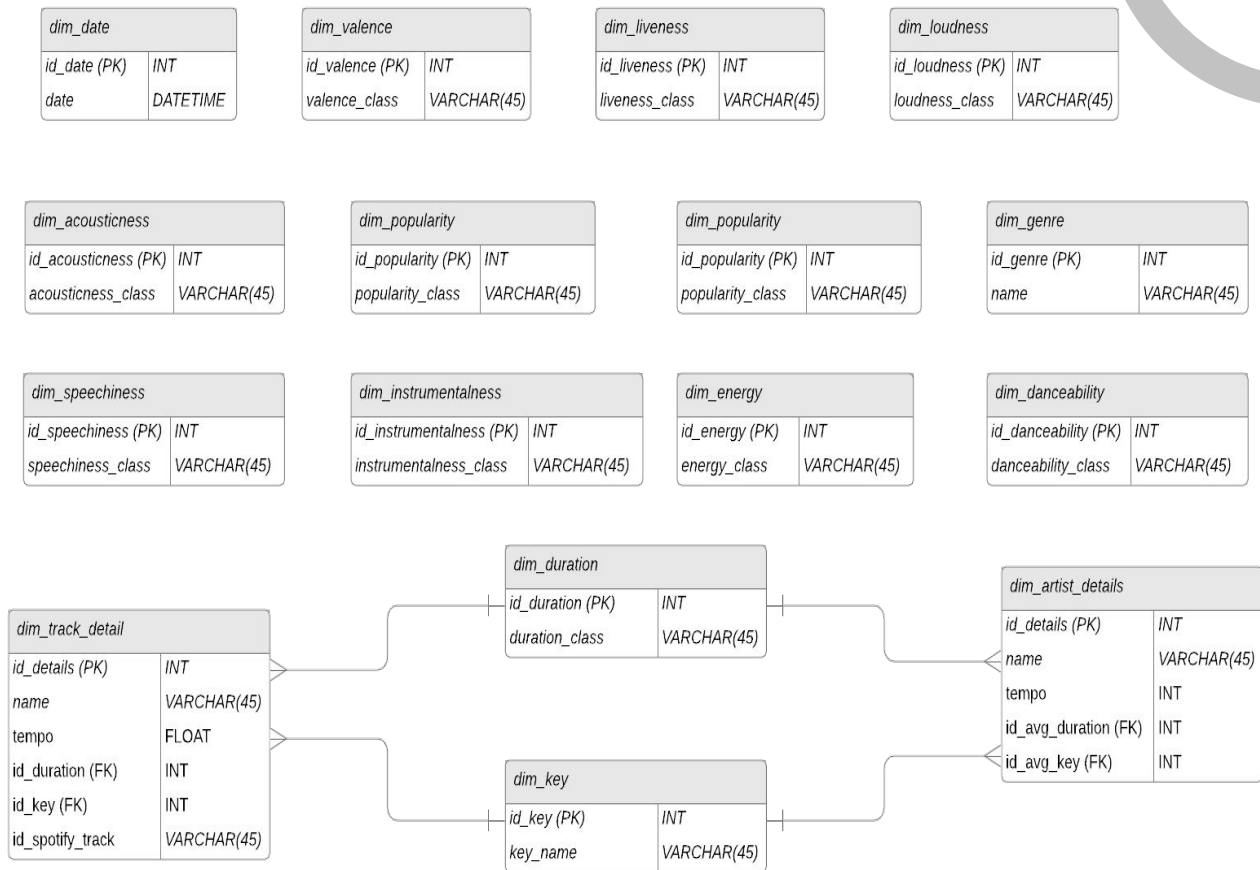


Fig.1- Dimensões utilizadas para a BD

Tabelas de facto

fact_track	
id_track (PK)	INT
mode	TINYINT
explicit	TINYINT
id_release_date (FK)	INT
id_valence (FK)	INT
id_details (FK)	INT
id_acousticness (FK)	INT
id_danceability (FK)	INT
id_energy (FK)	INT
id_instrumentalness (FK)	INT
id_liveness (FK)	INT
id_loudness (FK)	INT
id_popularity (FK)	INT
id_speechiness (FK)	INT

fact_artist	
id_artist (PK)	INT
mode	TINYINT
id_valence (FK)	INT
id_details (FK)	INT
id_acousticness (FK)	INT
id_danceability (FK)	INT
id_energy (FK)	INT
id_instrumentalness (FK)	INT
id_liveness (FK)	INT
id_loudness (FK)	INT
id_popularity (FK)	INT
id_speechiness (FK)	INT

Fig.2- Tabelas de facto utilizadas para a BD

Modelo Dimensional

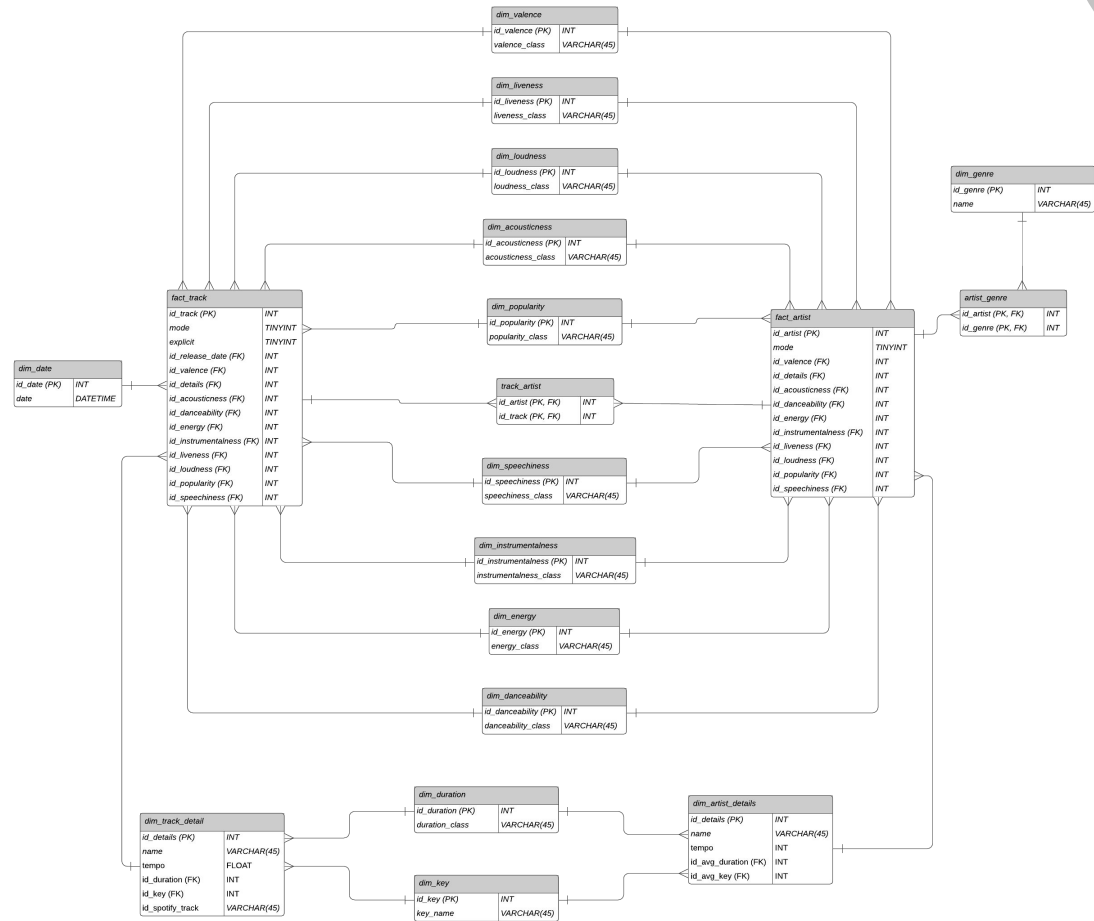


Fig.3- Modelo lógico da BD



03

Processos de ETL

Extração

- Optou-se por usar o “Import Wizard” diretamente de um ficheiro CSV, no nosso caso tínhamos três CSV’s distintos, um com os dados das músicas, um com os dados dos artistas e por último um com o mapeamento dos “*encondings*” das notas musicais.
- Estes *imports* foram feitos para um schema temporário, criado para receber e tratar estes dados, a nossa *Staging Area*.
- Seguiu-se um *dump* da nossa estrutura de dados da *Staging Area* para um script SQL para facilitar e acelerar o processo.

Transformação

- Para se fazer uma classificação das métricas em gamas de valores foram criadas funções para essas mesmas classificações.

```
DELIMITER |
CREATE FUNCTION duration_classification (duration double)
RETURNS int
DETERMINISTIC
BEGIN
    DECLARE classification int;
    CASE
        WHEN duration <= 60000 THEN SET classification = 1;
        WHEN duration > 60000 AND duration <= 120000 THEN SET classification = 2;
        WHEN duration > 120000 AND duration <= 180000 THEN SET classification = 3;
        WHEN duration > 180000 AND duration <= 240000 THEN SET classification = 4;
        WHEN duration > 240000 AND duration <= 300000 THEN SET classification = 5;
        WHEN duration > 300000 AND duration <= 360000 THEN SET classification = 6;
        ELSE SET classification = 7;
    END CASE;
    RETURN (classification);
END;
|
DELIMITER ;
```

Fig.4 - Função de classificação do tempo de duração de uma música

Transformação

```
DELIMITER |
CREATE FUNCTION valence_classification (valence double)
RETURNS int
DETERMINISTIC
BEGIN
    DECLARE classification int;
    CASE
        WHEN valence >= 0 AND valence <= 0.2 THEN SET classification = 1;
        WHEN valence > 0.2 AND valence <= 0.4 THEN SET classification = 2;
        WHEN valence > 0.4 AND valence <= 0.6 THEN SET classification = 3;
        WHEN valence > 0.6 AND valence <= 0.8 THEN SET classification = 4;
        WHEN valence > 0.8 AND valence <= 1 THEN SET classification = 5;
    END CASE;
    RETURN (classification);
END;
|
DELIMITER ;
```

Fig.5 - Função de classificação da propriedade valence

Transformação

- Para tratar dos valores presentes na lista de artistas e géneros criou-se procedures para popular as respetivas dimensões corretas

```
DELIMITER |
CREATE PROCEDURE populate_dim_genre (bound VARCHAR(255))
BEGIN

DECLARE id INT DEFAULT 0;
DECLARE value TEXT;
DECLARE occurrence INT DEFAULT 0;
DECLARE i INT DEFAULT 0;
DECLARE COUNT INT;
DECLARE splitted_value VARCHAR(255);
DECLARE done INT DEFAULT 0;
DECLARE cur1 CURSOR FOR SELECT distinct
    SUBSTR(genres,INSTR(genres,',' )+1,INSTR(genres,',' ) -(1+INSTR(genres,',' )))
    FROM spotify_staging.data_w_genres
    WHERE genres != ',';

DECLARE CONTINUE HANDLER FOR NOT FOUND SET done = 1;

OPEN cur1;
read_loop: LOOP
    FETCH cur1 INTO value;
    IF done THEN
        LEAVE read_loop;
    END IF;

    SET occurrence = (SELECT LENGTH(value) - LENGTH(REPLACE(value, bound, '')) + 1);
    SET i=1;
    WHILE i <= occurrence DO
        SET splitted_value = (SELECT LTRIM(REPLACE(SUBSTRING(SUBSTRING_INDEX(value, bound, i),
            LENGTH(SUBSTRING_INDEX(value, bound, i - 1)) + 1), ',' , ''));
        SET COUNT = (SELECT COUNT(*) FROM dim_genre WHERE name=splitted_value);
        IF COUNT = 0 THEN
            INSERT INTO dim_genre (name) VALUES (splitted_value);
        END IF;
        SET i = i + 1;
    END WHILE;
END LOOP;
CLOSE cur1;
END;
|
DELIMITER ;
```

Fig.6 - Procedure para separar os valores presentes na lista de géneros

Carregamento

- Nesta fase, o nosso objetivo era transferir os dados da nossa *staging area*, criada na fase de extração, para o nosso *data warehouse*.
- Para esse objetivo fez-se o carregamento dos dados das dimensões independentes, de seguida preencheu-se as dimensões dependentes de outras dimensões, depois o carregamento das tabelas de facto e por fim a população das tabelas que faziam uma relação de músicas com artistas e vice-versa.



04

Sistema de BI

Interrogações propostas

- Artistas com maior número de músicas;
- Anos mais populares;
- Artistas mais populares;
- Número de artistas populares por década;
- Emoções apresentadas nas músicas de 1929 vs 2020;
- Artistas pop populares por duração;
- Géneros de músicas mais populares;
- Rácio entre qualidade de som e popularidade;
- Rácio entre energia e nível acústico;
- Relação entre a emoção e a sua capacidade de dançar;
- Músicas rock populares e a sua energia.

Artistas com maior número de músicas

Artists with the most number of musics



Fig.7- Artistas com maior número de músicas

Anos mais populares

Average popularity by year

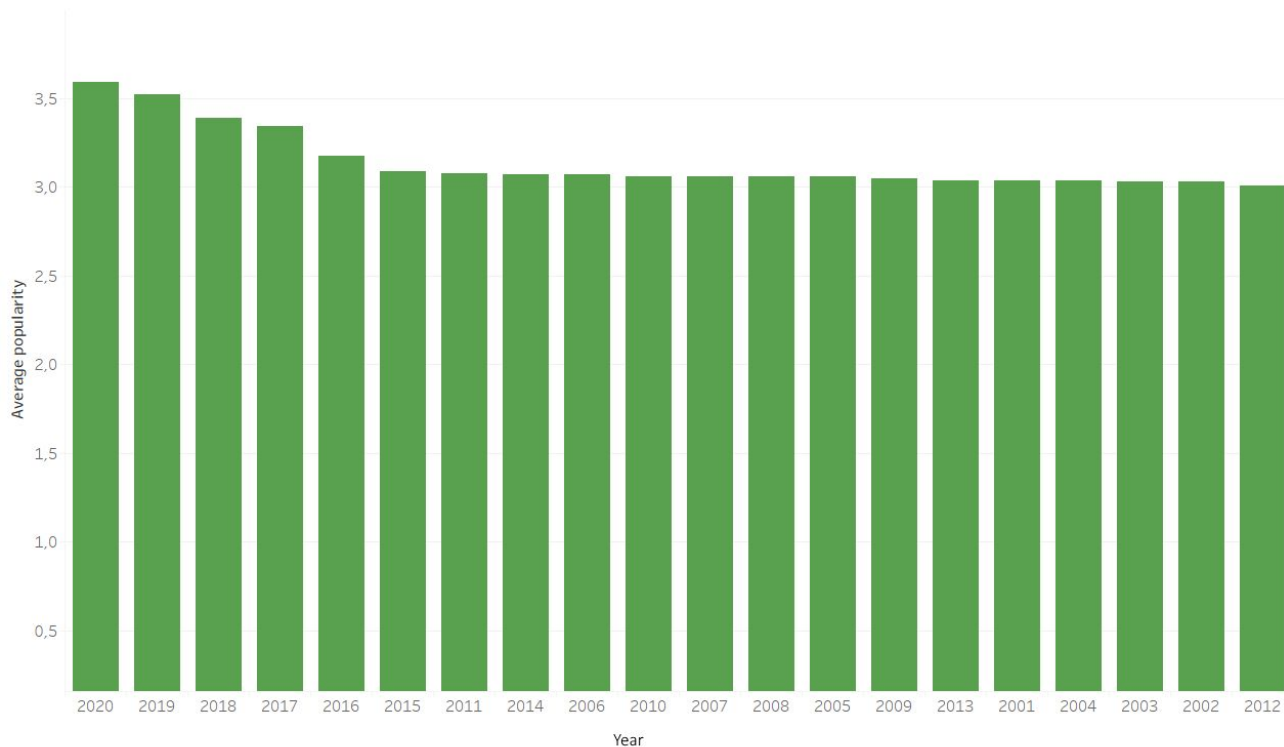


Fig.8- Média da popularidade em cada ano

Artistas mais populares

Top 9 artists by popularity

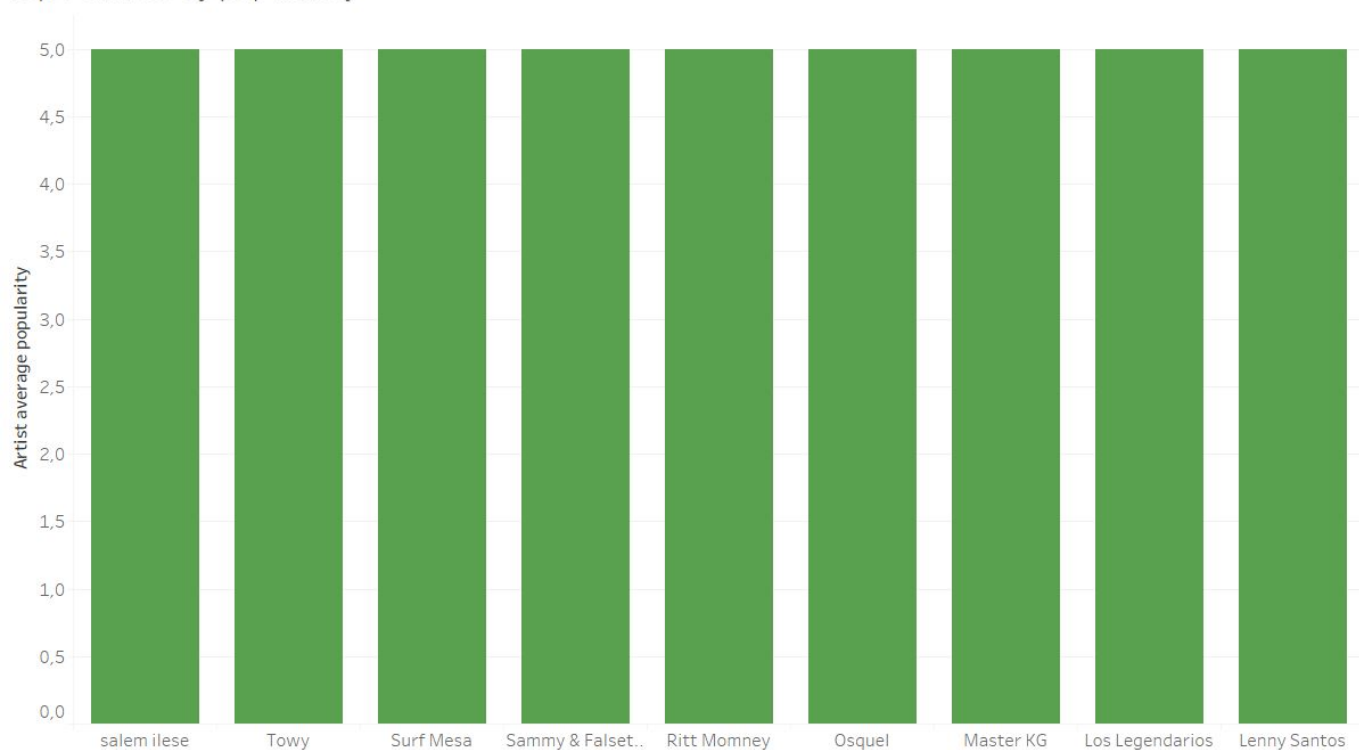


Fig.9- Os 9 artistas mais populares

Número de artistas populares por década

Number of artists by popularity each decade

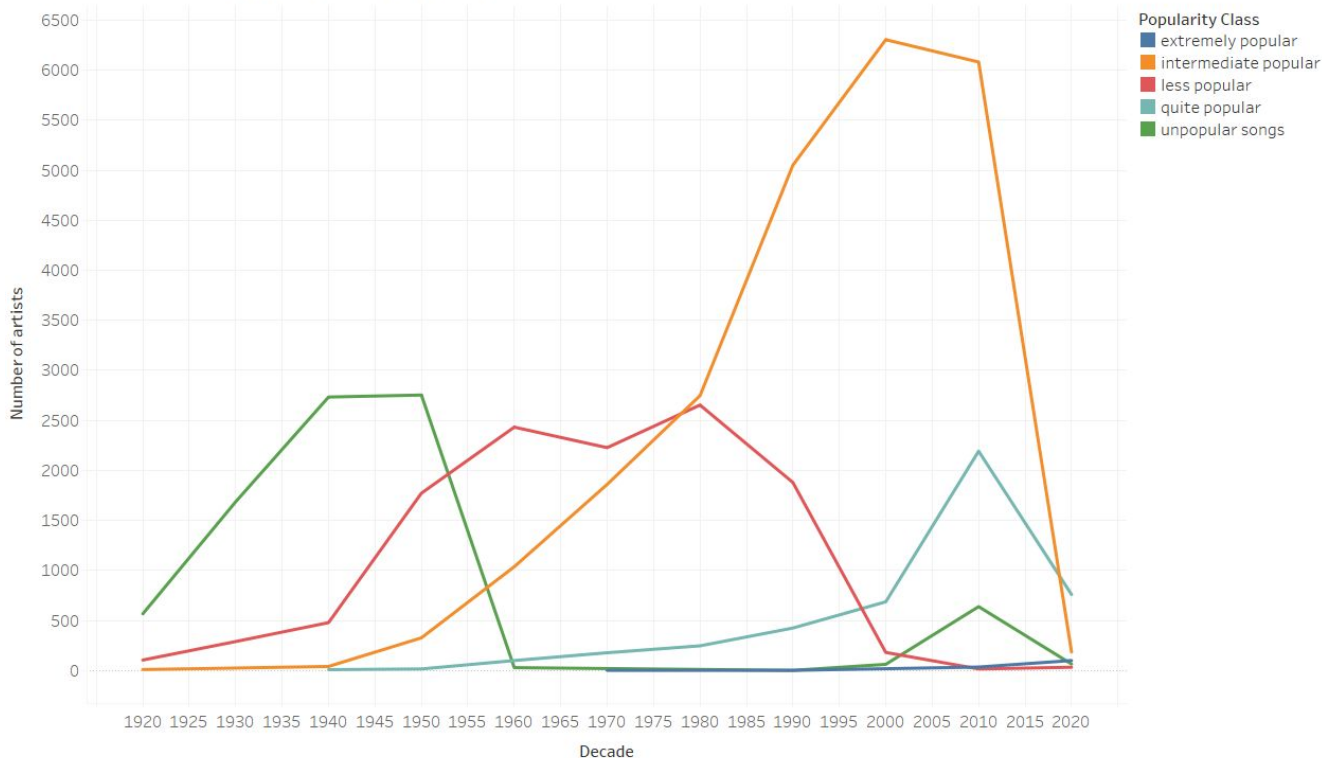
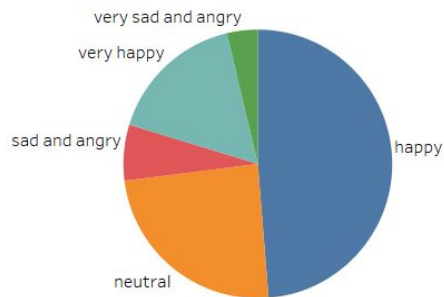


Fig.10- Número de artistas populares por décadas

1929 vs 2020

Distribution of Valence (Happiness) in 1929



Distribution of Valence (Happiness) in 2020



Valence Class

happy

neutral

sad and

very happy

very sad and angry

Fig.11- Comparação dos valores de felicidade de uma música no ano 1929 com 2020

Artistas pop populares por duração

Pop artists popularity by song duration

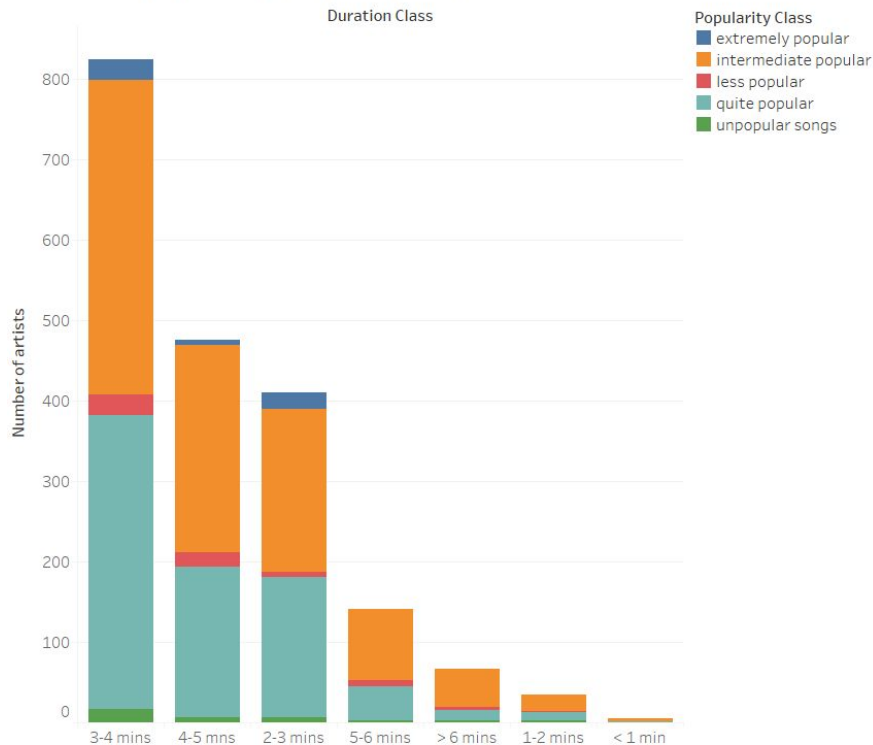


Fig.12- Popularidade dos artistas pop em relação à duração das suas músicas

Gêneros de música mais populares

Top 3 genres by popularity

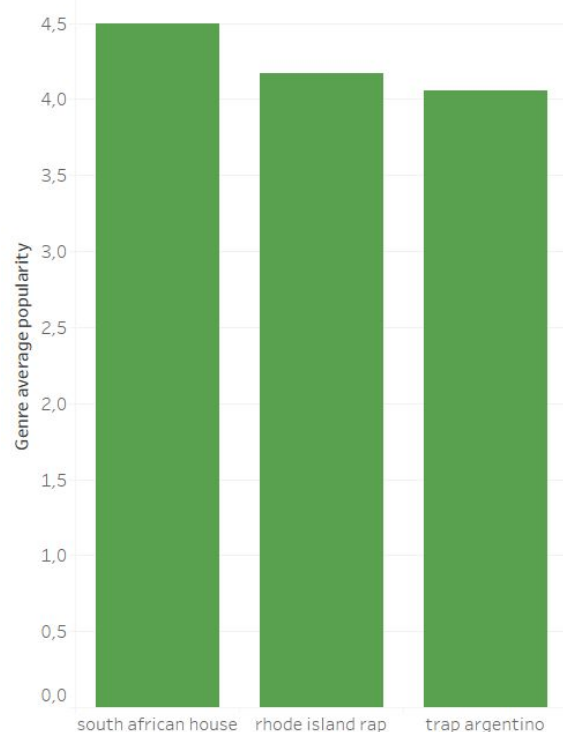


Fig.13- Os 3 gêneros de música com maior popularidade

Rácio entre qualidade de som e popularidade

Ratio between loudness and popularity

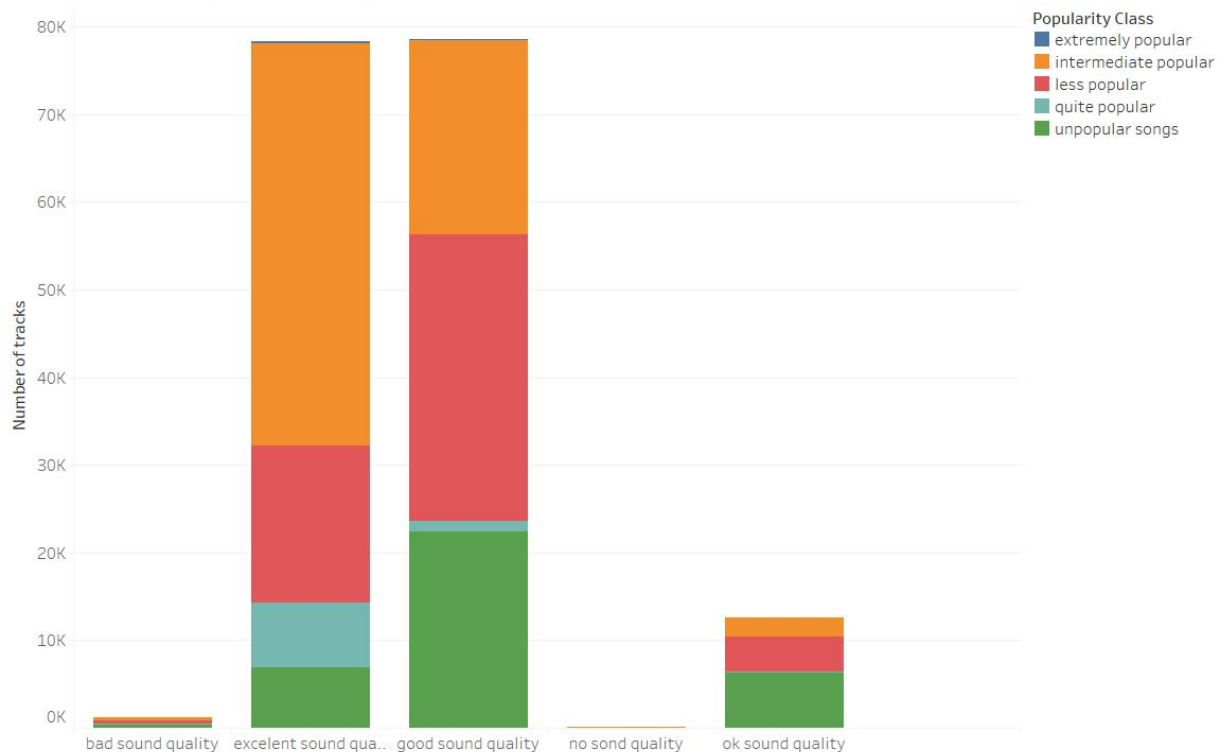


Fig.14- Popularidade das músicas em relação à qualidade de som

Rácio entre energia e nível acústico

Ratio between energy and accousticness

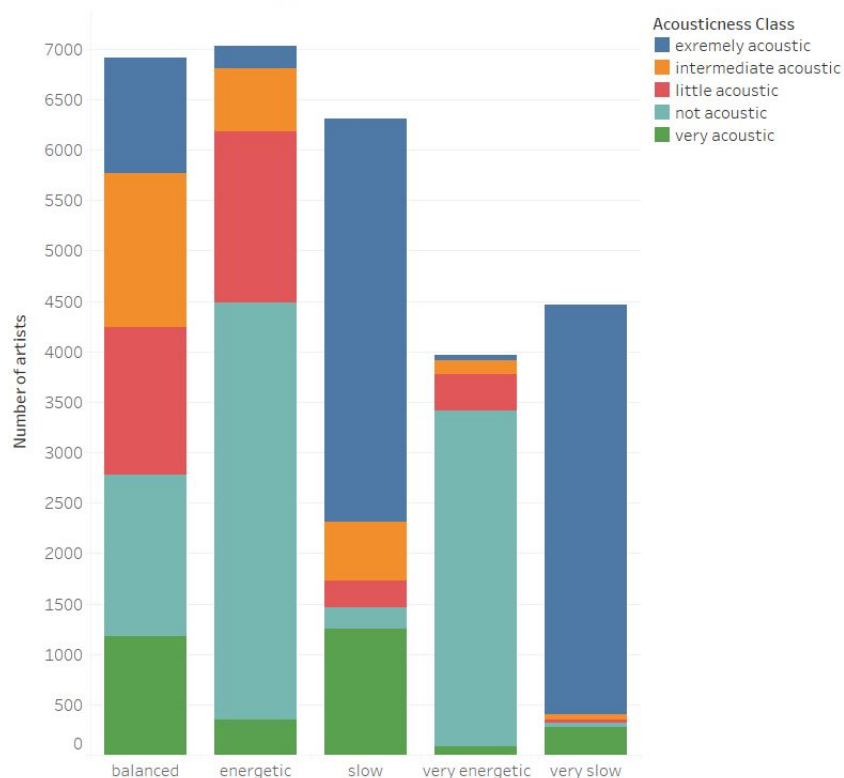
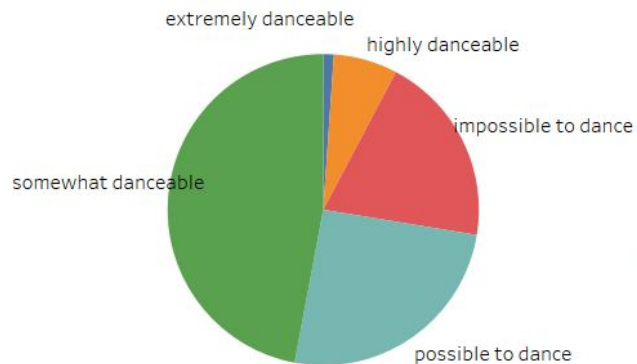


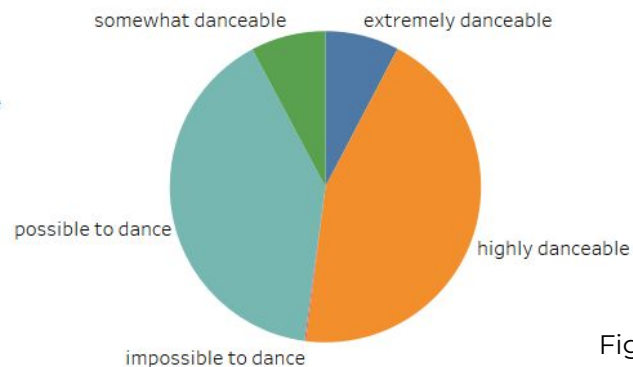
Fig.15- relação entre a energia e nível acústico de uma música

Relação entre a Emoção e a Capacidade de Dançar

Relation between valence and
dancibility (very sad and angry)



Relation between valence and
dancibility (happy)



Danceability Class



extremely danceable



highly danceable



impossible to dance



possible to dance



somewhat danceable

Fig.16- Relação entre a
emoção e capacidade de
dançar de uma música

Músicas Rock populares e a sua energia

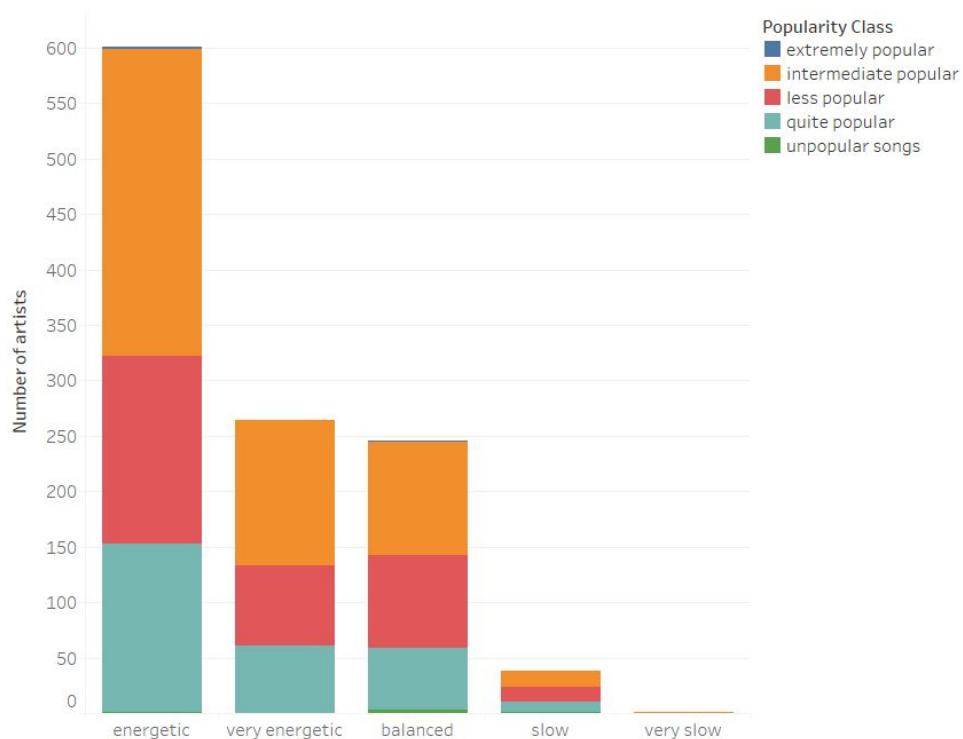


Fig.17- Relação entre a energia presente nas músicas de Rock e a sua popularidade

Conclusão

- Possibilidade de compreender melhor a arquitetura de um data warehouse.
- Conseguir identificar um dataset adequado e a partir do mesmo fazer uma modelação dimensional.
- Capacidade de fazer a importação dos dados para o nosso data warehouse através de processos de ETL.
- Possibilidade de exploração da ferramenta Tableau e a capacidade de fazer uma análise dos dados do data warehouse.