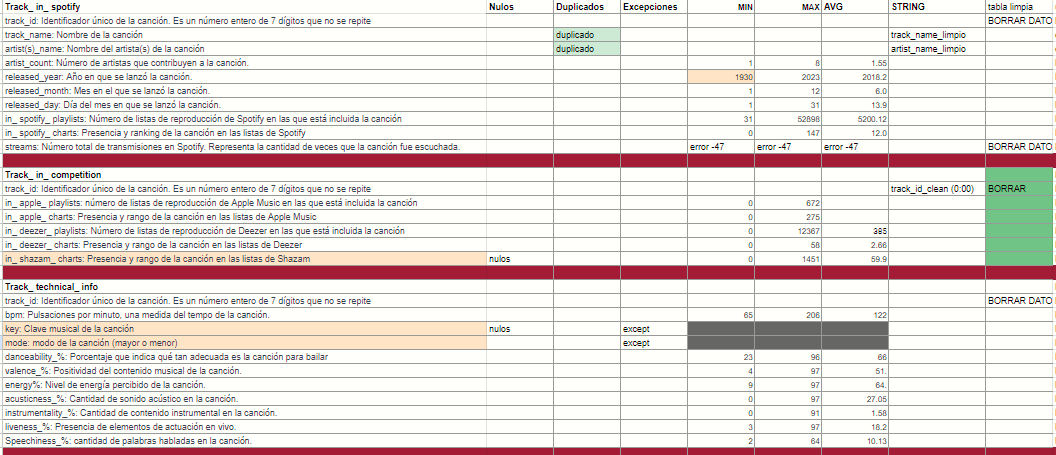
**Documentación**

1 PASO

Identificar la limpieza necesaria por cada tabla:

en total tenemos tres tablas: SPOTIFY, COMPETITION Y TECHNICAL.



Una vez identificadas las necesidades se comienza con la unión del código por cada tabla.

**COMPETITION**

* ~~borrar de track\_id: 0:00~~
* ~~borrar el los datos del “stream” que se eliminará track\_id : 4061483 (~~

| Love Grows (Where My Rosemary Goes) | Edison Lighthouse |
| --- | --- |

* eliminar duplicados

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.competition`

WHERE track\_id = '0:00'\*/

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.competition`

WHERE track\_id = '4061483'\*/

SELECT \*,

IFNULL(SAFE\_CAST(in\_shazam\_charts AS INT64),0) AS in\_shazam\_charts\_clean

FROM

`proyecto-hipotesis-404820.dataset2.competition`

WHERE

track\_id!= "3814670"

AND track\_id != "7173596"

AND track\_id != "4586215"

AND track\_id != "1119309"

AND track\_id != "5080031"

AND track\_id != "8173823"

AND track\_id != "4967469"

AND track\_id != "5675634"

**SPOTIFY**

* ~~se identificaron duplicados~~
* ~~se borró de track\_id el 0:00~~
* ~~se borró el track\_id: 4061483 (dato de streams malo~~
* ~~Se quitaron carácteres especiales de track\_name y artist\_name~~
* ~~se ajustaron datos nulos a “0”de shazam\_charts~~

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.spotify`

WHERE track\_id = '0:00'\*/

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.spotify`

WHERE track\_id = '4061483'\*/

/\*se uso delete para eliminar el dato del streams que tenía letras\*/

SELECT

\*,

ROW\_NUMBER() OVER (PARTITION BY track\_name\_clean, artist\_name\_clean ORDER BY streams\_clean) AS duplicados

FROM (

SELECT

\*,

REGEXP\_REPLACE(track\_name, r'[^a-zA-Z0-9-]', ' ') AS track\_name\_clean,

REGEXP\_REPLACE(artist\_s\_\_name, r'[^a-zA-Z0-9-]', ' ') AS artist\_name\_clean,

IFNULL(SAFE\_CAST(streams AS INT64), 0) AS streams\_clean,

DATE(CONCAT(CAST(released\_year AS STRING), '-', LPAD(CAST(released\_month AS STRING), 2, '0'), '-', LPAD(CAST(released\_day AS STRING), 2, '0'))) AS date\_new,

FROM

`proyecto-hipotesis-404820.dataset2.spotify`

WHERE

streams NOT LIKE "BMP%"

AND track\_id!= "3814670"

AND track\_id != "7173596"

AND track\_id != "4586215"

AND track\_id != "1119309"

AND track\_id != "5080031"

AND track\_id != "8173823"

AND track\_id != "4967469"

AND track\_id != "5675634"

);

/\*en total son 943 datos los que quedan, se eliminaron 8 de duplicaods, 1 de streams y 1 más de track id\*/

**TECHNICAL**

* ~~borrar id\_track: 0:00~~
* ~~borrar el dato de streams con id\_track: 4061483~~
* ~~Sacar key y mode como variables (son muchos nulos)~~

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.technical`

WHERE track\_id = '0:00'\*/

/\*DELETE FROM `proyecto-hipotesis-404820.dataset2.technical`

WHERE track\_id = '4061483'\*/

SELECT \*

EXCEPT (key, mode)

FROM `proyecto-hipotesis-404820.dataset2.technical`

WHERE

track\_id!= "3814670"

AND track\_id != "7173596"

AND track\_id != "4586215"

AND track\_id != "1119309"

AND track\_id != "5080031"

AND track\_id != "8173823"

AND track\_id != "4967469"

AND track\_id != "5675634"

/\*estoy eliminando los duplicados\*/

**CREAR NUEVA VARIABLE**

FECHA CONCATENAR

**SPOTIFY**

SELECT

CONCAT(

CAST(released\_year AS STRING),

'-',

LPAD(CAST(released\_month AS STRING), 2, '0'), -- Agrega un cero al mes si es necesario

'-',

LPAD(CAST(released\_day AS STRING), 2, '0') -- Agrega un cero al día si es necesario

) AS date\_new,

DATE(CONCAT(CAST(released\_year AS STRING), '-', LPAD(CAST(released\_month AS STRING), 2, '0'), '-', LPAD(CAST(released\_day AS STRING), 2, '0'))) AS date\_new,

FROM `proyecto-hipotesis-404820.dataset2.spotify`

Crear una subconsulta para unir las tablas limpias donde se vea la tabla con las variables limpias o columnas ya bien

ejemplo

select \* except

streams, (las que no son limpias)

creo que se puede generar una subconsulta dentro de la misma consulta y tabla

**CREACIÓN DE NUEVA TABLA-PRUEBA**

**SPOTIFY**

CREATE TABLE `proyecto-hipotesis-404820.dataset2.spotify\_clean`

AS

SELECT \* EXCEPT (track\_name, artist\_s\_\_name, streams)

FROM `proyecto-hipotesis-404820.dataset2.spotify\_cleaned`

**TECHNICAL**

CREATE TABLE `proyecto-hipotesis-404820.dataset2.technical\_clean`

AS

SELECT \*

FROM `proyecto-hipotesis-404820.dataset2.technical\_cleaned`

/\*technical\_clean es mi consulta y technical\_cleaned es mi vista\*/

/\*spotify\_clean es el nombre de la consulta, el spotify\_clean\_clean es el nombre de mi vista y puedo hacer excepciones de que columnas no quiero que me traiga\*/

**COMPETITION**

CREATE TABLE `proyecto-hipotesis-404820.dataset2.competition\_clean`

AS SELECT

\* EXCEPT (in\_shazam\_charts)

FROM `proyecto-hipotesis-404820.dataset2.competition\_cleaned`

**UNIR TABLAS**

TABLA 2

**CONJUNTO DE TABLA DE SPOTIFY Y TECHNICAL**

SELECT \*

FROM

`dataset2.spotify\_clean` AS spotify\_tabla1

LEFT JOIN

`dataset2.technical\_clean` AS technical\_tabla2

ON

spotify\_tabla1.track\_id = technical\_tabla2.track\_id

/\*dos tablas unidas primero la de spotify y la de technical\*/

**TABLA 3**

CONJUNTO TABLA 2 CON COMPETITION

SELECT \*

EXCEPT (track\_id\_1)

FROM

`dataset2.tabla\_spotify\_y\_tech` AS tabla\_spot\_tech

LEFT JOIN

`dataset2.competition\_clean` AS comp\_tabla\_3

ON

tabla\_spot\_tech.track\_id = comp\_tabla\_3.track\_id

/\*antes de crear una segunda tabla que contenga las tres tablas, debo crear la tabla y no traerla desde una vista o consulta \*/

El problema con este código para unir las tablas es que se duplican las columnas de track\_id, y al momento de borrarlas se pierde información, es por ello que es necesario renombrar las columnas con AS para que podamos eliminarlas o ponerle EXCEPT y que logre identificar una a una cada tabla sin ambigüedad.

NUEVA VARIABLE

CREATE TABLE `proyecto-hipotesis-404820.dataset2.artist\_song\_counts` AS

WITH artist\_solos AS(

SELECT

artist\_name\_clean,

COUNT(track\_id) AS total\_canciones

FROM `proyecto-hipotesis-404820.dataset2.spotify\_cleaned`

WHERE

artist\_count = 1

GROUP BY

artist\_name\_clean

)

SELECT

artist\_name\_clean,

total\_canciones

FROM

artist\_solos;

Este código cuenta el número de canciones que cada artista tiene de manera solitaria, agrupando por nombre de artista y contando donde diga el número de artistas que colaboran en la canción y diga 1, al final cuenta de cada artista.