Find Duplicate Values in a Table

HAVING count (*) >1

id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Context

Challenge

BigQuery

Method

Raw Code

Result

Context

Description

Duplicate rows have been created through human error or uncleaned data from external sources. Generally, Data Analysts role is to find those values and remove them.

Sometimes though duplicate values can be useful when new data is associated to a given dimension. In particular, when values are not overwritten.

NIF	Value.1	Value.2	Value.3
A123	abc	abc	abc
B456	“	“	“
C789	“	“	“
(n)	“	“	“
A123	abc	def	hij

Challenge

Identify Entities that have received funds in different years

By querying a public record database we'll identify those entities that appear in the database more than one time.

Then we'll fetch them along with its dimensions to facilitate a wider understanding of their traits.

id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Why BigQuery?

Description

When it comes to Analytics, reliance on a single source of data is not enough to make strategic decisions. Any online business uses different digital tools that in turn generate useful data.

To have an overview of the entire digital ecosystem and get meaningful insights, it is advisable to count on data warehouse solutions with access to any of these sources.

BigQuery is a good example as it naturally integrates with any of its products such as spreadsheets, GA4, Google Ads, Data Studio, etc.

Method

Step 1 - Identify Duplicate Criteria

The first step is to define your criteria for a duplicate row. Do you need a combination of two columns to be unique together, or are you simply searching for duplicates in a single column?

In this example, we are searching for duplicates in a single column in our table: **NIF** (entity_id)

NIF	nom_entitat	nom_projecte
G63306914	PLATONIQ, SISTEMA CULTURAL	Data Camp
G66025784	ASSOCIACIO CULTURAL ESPAI ERRE	ROTOR FAB
G65338154	ASSOCIACIO PANORAMA 180	Subvenció TIC CCworld
G65744138	ASSOCIACIÓ REBOBINART	APP Wallspot
J66647025	MELEZ SCP	Subvenció Mélez
G64643547	IDENSITAT ASSOCIACIO D'ART CONTEN	Mediterrànies_Idensitat
G66427162	CFD BARCELONA - CENTRE DE FOTOG	TICs 2016
J62241583	TELEDUCA.EDUCACIO I COMUNICACIO	Cinema amb ulls d'infant
G64319072	ASSOCIACIO KONICLAB, CREACIO CON	Dispositiu en Xarxa obert al públic, per a Jorr
G63832786	FUND PRIV CIUTADANIA MULTICULTUR	Ciudad migrante
G66262635	ASSOCIACIÓ EDUCATIVA I CULTURAL S	Subvenció
G60939956	ASSOCIACIO PER A JOVES TEB	PFI - projecte de vídeo porter
G64031586	HABITUAL VIDEO TEAM	Subvencions per a inversions en tecnologies
G64219454	ASSOCIACIO RUIDO PHOTO	Campanya de sensibilització amb video 360
G59809665	TANTAGORA SERVEIS CULTURALS	Sorolls
G64267065	ASSOCIACIO HAMACA	Renovació equip tècnic Hamaca
G64210768	ASSOCIACIÓ LIVEMEDIA	Enfortiment Activitats Associació Livemedia
B66164450	LOVE STREAMS SL	Screenly Non Theatrical
B65236689	POLYPLICITY SOCIEDAD LIMITADA	FabCAfe School

Method

Step 2 - Verify that Duplicates Exist

The next query verifies whether duplicates do indeed exist in the table. If so, rows will be returned.

We've counted how many times each entity has been recorded more than 1 time.

```
SELECT NIF, nom_entitat, COUNT(*) AS duplicate_times
FROM `xxxxxtable_A_xxxxxxx`
GROUP BY NIF,nom_entitat
HAVING COUNT(*) > 1
```

NIF	nom_entitat	duplicate_times
G64643547	IDENSITAT ASSOCIACIO D'ART CONTEMPORANI	3
G66427162	CFD BARCELONA FOTOGRAFIA I MITJANS DOCUMENTALS	3
G59809665	TANTAGORA SERVEIS CULTURALS	2
B65236689	POLYPLICITY SOCIEDAD LIMITADA	2
F08310013	COOP. PROMOTORA MEDIOS AUDIOVISUALES	2
F66864778	ZUMZEIG CINECOOPERATIVA	2
G65744138	ASSOCIACIÓ REBOBINART	2
B65236689	POLYPLICITY SOCIEDAD LIMITADA	2
****3	ALBERT ARGILÉS LLORENS	2
B60758109	TANTARANTANA TEATRE, SL	4

Query

Step 3 - SQL Code

Now, we want to return the entire record for each duplicate row. To accomplish this, we'll need to select the entire table and join that to our duplicate rows.

Because we're joining the table to itself, it's necessary to use aliases (here, we're using A and B) to label the two versions.

1 → Select everything from table A

2 → Join the table with the duplicates rows

3 → Now filter (Having count) if *NIF* appears more than one

```
1  SELECT A.*
   FROM   `xxxxxtable_A_xxxxxxx` A

2  JOIN (SELECT NIF, COUNT(*)
   FROM  `xxxxxtable_A_xxxxxxx`
   GROUP BY NIF

3  HAVING count(*) >1 ) B
   USING (NIF)
   ORDER BY NIF, periode
```

Result

Organized Table with Duplicate Entities

Now we know how much funds each entity has received from in different years.

NIF	modalitat	nom_entitat	nom_projecte	estat	import_otorgat	punts_obtinguts	periode
4****793P	A	C.S.B	BESTIARI DIGITAL - ESCANEJAT 3D DEL BESTIARI FESTIU DE LA CIUTAT	denegada	0	3,33	2018
4****793P	A	C.S.B	BESTIARI DIGITAL - ESCANEJAT 3D DEL BESTIARI FESTIU DE LA CIUTAT	denegada	0	0	2018
B60758109	A	TANTARANTANA TEATRE, SL	Tantared	denegada	0	no sabem	2016
B60758109	B	TANTARANTANA TEATRE, SL	Inversions Tic Tanta	acceptada	6630	9,5	2017
B60758109	B	TANTARANTANA TEATRE, SL	Projectes Escènics de creativitat digital	acceptada	6806	5,5	2018
B60758109	B	TANTARANTANA TEATRE, SL	TELÓ DIGITAL	denegada	0	0	2018
B61013520	B	AMBIT GALERIA D'ART, SL	Subvenció TIC	denegada	0	no sabem	2016
B61013520	B	AMBIT GALERIA D'ART, SL	Inversió en infraestructura tecnològica	acceptada	1948	5,5	2017
B63783583	B	INTELLIGENT CONSULTING, SL	Tecnologies Espacials i Ciutadans	denegada	0	no sabem	2016
B63783583	B	INTELLIGENT CONSULTING, SL	Tecnologies Espacials i Ciutadans	denegada	0	3,5	2017
B65236689	B	POLYPLICITY SOCIEDAD LIMITADA	FabCAfe School	acceptada	1413	no sabem	2016
B65236689	B	POLYPLICITY SOCIEDAD LIMITADA	Polyplicity People Fund People	acceptada	5390	no sabem	2016
B65236689	A	POLYPLICITY SOCIEDAD LIMITADA	Assemblers	denegada	0	no sabem	2016
B65236689	B	POLYPLICITY SOCIEDAD LIMITADA	Fabcafé Laser	denegada	0	no sabem	2016

Made by

see website