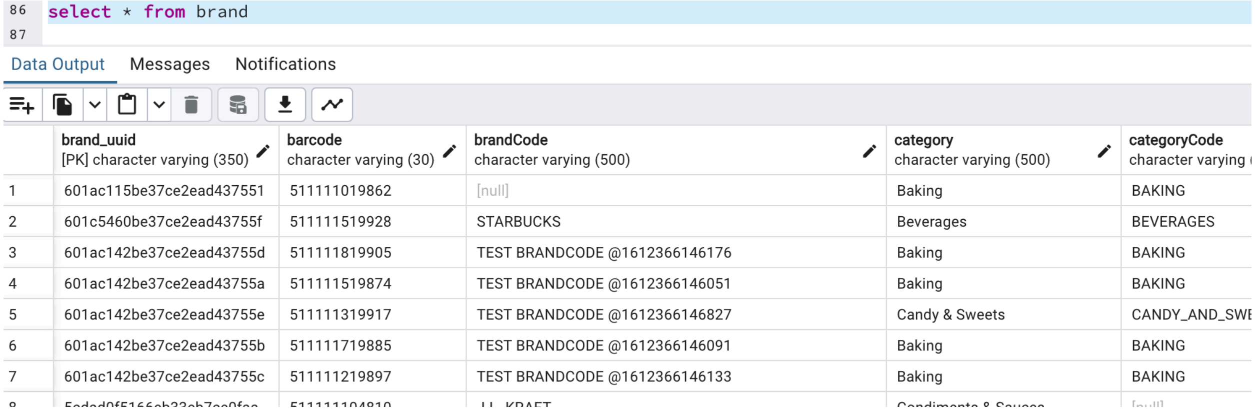
Brands data quality checks –

select * from brand



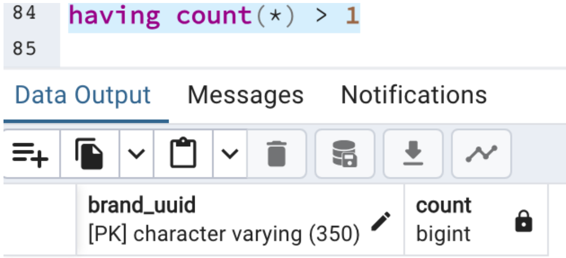
```
select * from brand
```

	brand_uuid [PK] character varying (350)	barcode character varying (30)	brandCode character varying (500)	category character varying (500)	categoryCode character varying (
1	601ac115be37ce2ead437551	511111019862	[null]	Baking	BAKING
2	601c5460be37ce2ead43755f	511111519928	STARBUCKS	Beverages	BEVERAGES
3	601ac142be37ce2ead43755d	511111819905	TEST BRANDCODE @1612366146176	Baking	BAKING
4	601ac142be37ce2ead43755a	511111519874	TEST BRANDCODE @1612366146051	Baking	BAKING
5	601ac142be37ce2ead43755e	511111319917	TEST BRANDCODE @1612366146827	Candy & Sweets	CANDY_AND_SWE
6	601ac142be37ce2ead43755b	511111719885	TEST BRANDCODE @1612366146091	Baking	BAKING
7	601ac142be37ce2ead43755c	511111219897	TEST BRANDCODE @1612366146133	Baking	BAKING
8	5cdad0f5166eb33eb7ce0faa	511111104810	J.L. KRAFT	Condiments & Sauces	[null]

select brand_uuid, count(*) from brand

group by brand_uuid

having count(*) > 1



```
having count(*) > 1
```

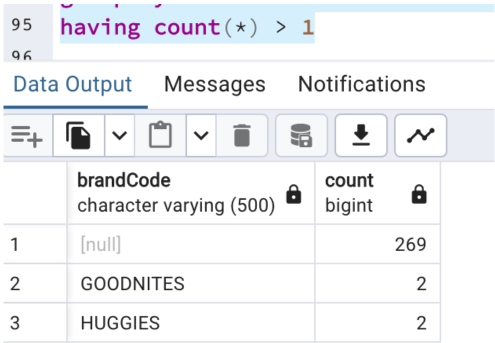
	brand_uuid [PK] character varying (350)	count bigint

select "brandCode", count(*)

from brand

group by "brandCode"

having count(*) > 1

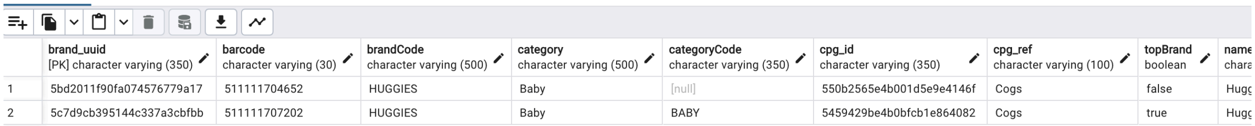


```
having count(*) > 1
```

	brandCode character varying (500)	count bigint
1	[null]	269
2	GOODNITES	2
3	HUGGIES	2

Although just 2 duplicates, but a lot of NULLS for brandcode, depicting data quality issues and reducing reliability to use this field as a join. Let's dive further to inspect where the duplicates are there -

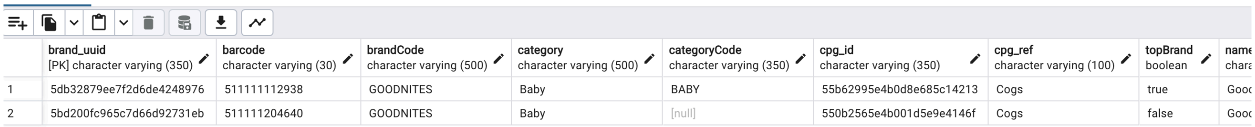
select * from brand where "brandCode" = 'HUGGIES'



	brand_uuid [PK] character varying (350)	barcode character varying (30)	brandCode character varying (500)	category character varying (500)	categoryCode character varying (350)	cpg_id character varying (350)	cpg_ref character varying (100)	topBrand boolean	name chara
1	5bd2011f90fa074576779a17	511111704652	HUGGIES	Baby	[null]	550b2565e4b001d5e9e4146f	Cogs	false	Hugg
2	5c7d9cb395144c337a3cbfbb	511111707202	HUGGIES	Baby	BABY	5459429be4b0bfcb1e864082	Cogs	true	Hugg

From my observation, I see that the barcode is different and also one of them has missing categoryCode. My intuition is that one of the records is an older one, which has not been removed from the table (since just 2 duplicates). However, I will have to ask leader in this regard for better decision making. One solution can be to transform this into SCD wherein we can then maintain history and use current Flag to mark the latest record and solve this issue.

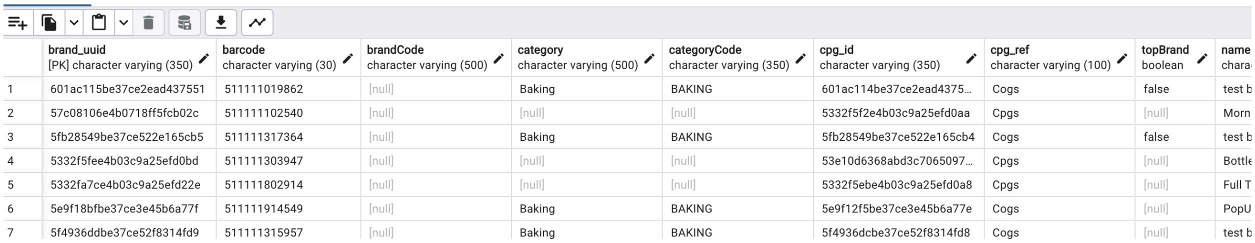
select * from brand where "brandCode" = 'GOODNITES'



	brand_uuid [PK] character varying (350)	barcode character varying (30)	brandCode character varying (500)	category character varying (500)	categoryCode character varying (350)	cpg_id character varying (350)	cpg_ref character varying (100)	topBrand boolean	name chara
1	5db32879ee7f2d6de4248976	511111112938	GOODNITES	Baby	BABY	55b62995e4b0d8e685c14213	Cogs	true	Gooc
2	5bd200fc965c7d66d92731eb	511111204640	GOODNITES	Baby	[null]	550b2565e4b001d5e9e4146f	Cogs	false	Gooc

Similar case for the other brandcode as well.

select * from brand where "brandCode" is NULL



	brand_uuid [PK] character varying (350)	barcode character varying (30)	brandCode character varying (500)	category character varying (500)	categoryCode character varying (350)	cpg_id character varying (350)	cpg_ref character varying (100)	topBrand boolean	name chara
1	601ac115be37ce2ead437551	511111019862	[null]	Baking	BAKING	601ac114be37ce2ead4375...	Cogs	false	test b
2	57c08106e4b0718ff5fcb02c	511111102540	[null]	[null]	[null]	5332f5f2e4b03c9a25efd0aa	Cpgs	[null]	Morn
3	5fb28549be37ce522e165cb5	511111317364	[null]	Baking	BAKING	5fb28549be37ce522e165cb4	Cogs	false	test b
4	5332f5fee4b03c9a25efd0bd	511111303947	[null]	[null]	[null]	53e10d6368abd3c7065097...	Cpgs	[null]	Bottle
5	5332fa7ce4b03c9a25efd22e	511111802914	[null]	[null]	[null]	5332f5ebe4b03c9a25efd0a8	Cpgs	[null]	Full T
6	5e9f18bfbe37ce3e45b6a77f	511111914549	[null]	Baking	BAKING	5e9f12f5be37ce3e45b6a77e	Cogs	[null]	PopU
7	5f4936ddbe37ce52f8314fd9	511111315957	[null]	Baking	BAKING	5f4936dcbe37ce52f8314fd8	Cogs	[null]	test b

On observing NULL Brandcode, I can observe topBrand as well as other fields and barcode being populated, thus confirming data quality issues with brandcode.

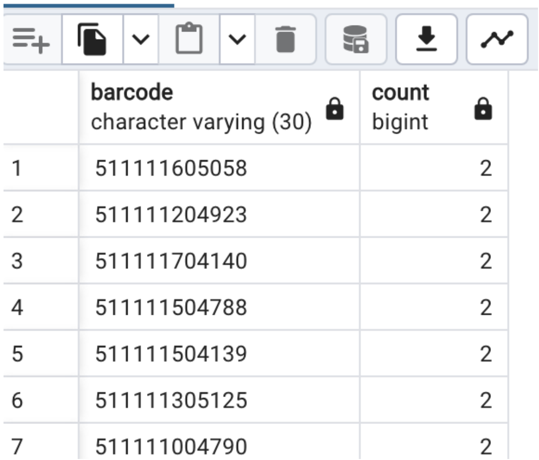
Now let's inspect barcode –

select barcode, count(*)

from brand

group by barcode

having count(*) > 1

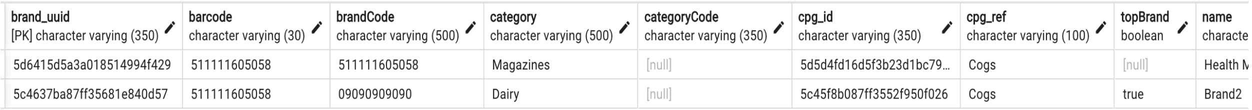


	barcode character varying (30)	count bigint
1	511111605058	2
2	511111204923	2
3	511111704140	2
4	511111504788	2
5	511111504139	2
6	511111305125	2
7	511111004790	2

We observe 7 duplicate barcodes as well but **no nulls**.

Diving into one of the barcodes –

select * from brand where barcode = '511111605058'



brand_uuid [PK] character varying (350)	barcode character varying (30)	brandCode character varying (500)	category character varying (500)	categoryCode character varying (350)	cpg_id character varying (350)	cpg_ref character varying (100)	topBrand boolean	name characte
5d6415d5a3a018514994f429	511111605058	511111605058	Magazines	[null]	5d5d4fd16d5f3b23d1bc79…	Cogs	[null]	Health M
5c4637ba87ff35681e840d57	511111605058	09090909090	Dairy	[null]	5c45f8b087ff3552f950f026	Cogs	true	Brand2

Here again I see data quality issues as **brandcode is also populated with barcode in one field** and these two are different 'names' itself. So we need to identify root cause for the issues here. Ideally we need to remove the duplicates, either by finding the source of truth table for this.

However, in order to proceed ahead with joins for data modeling, I have decided as of now based on NULL values and bad data in brandcode, and few duplicates, to use barcode to join with other tables in the data model.

```
ALTER TABLE brand
ADD CONSTRAINT pk_brand PRIMARY KEY (brand_uuid);
```