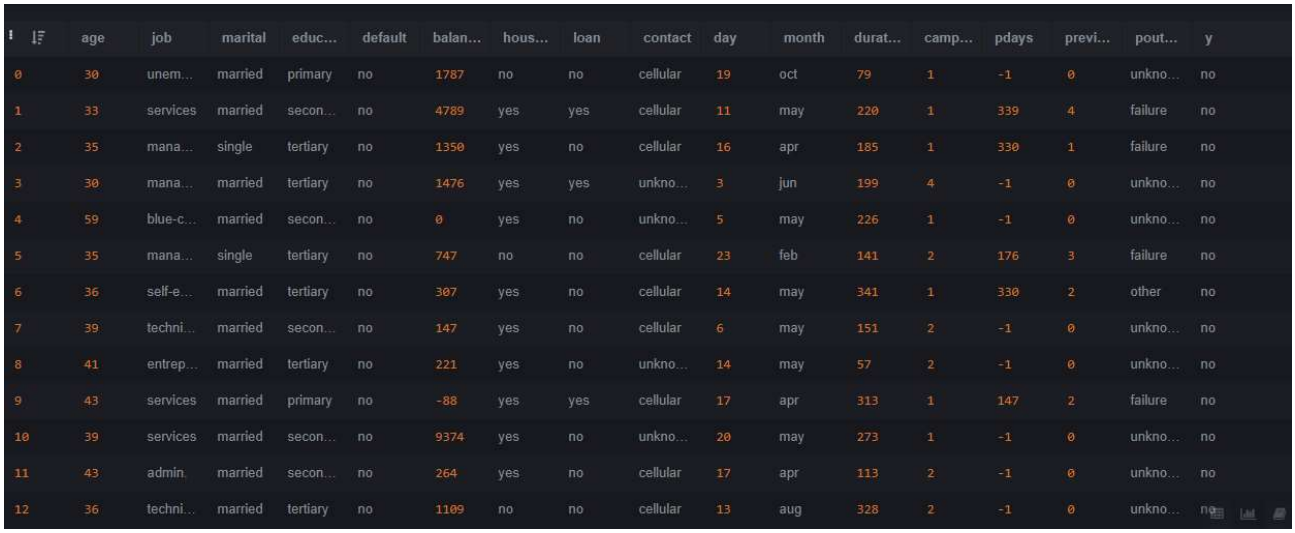
SQL: Resumen descriptivo de los datos

Vamos a visualizar nuestro Dataset *Bank Marketing Classification Task*, un conjunto de datos de una entidad bancaria portuguesa sobre una campaña de marketing para que los clientes se suscriban a un depósito a plazo, usando SQL.

Empezamos cargando los datos, llamados *bank_data*, y a continuación podemos seleccionar todos los datos usando:

```
1 SELECT *
2 FROM bank_data
3 ;
```

Con este código podemos observar el Dataset con todas las columnas:



	age	job	marital	educ...	default	balan...	hous...	loan	contact	day	month	durat...	camp...	pdays	previ...	pout...	y
0	30	unem...	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unkno...	no
1	33	services	married	secon...	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
2	35	mana...	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
3	30	mana...	married	tertiary	no	1476	yes	yes	unkno...	3	jun	199	4	-1	0	unkno...	no
4	59	blue-c...	married	secon...	no	0	yes	no	unkno...	5	may	226	1	-1	0	unkno...	no
5	35	mana...	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
6	36	self-e...	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
7	39	techni...	married	secon...	no	147	yes	no	cellular	6	may	151	2	-1	0	unkno...	no
8	41	entrep...	married	tertiary	no	221	yes	no	unkno...	14	may	57	2	-1	0	unkno...	no
9	43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
10	39	services	married	secon...	no	9374	yes	no	unkno...	20	may	273	1	-1	0	unkno...	no
11	43	admin.	married	secon...	no	264	yes	no	cellular	17	apr	113	2	-1	0	unkno...	no
12	36	techni...	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unkno...	no

Ahora, por ejemplo, podemos ordenar los datos por edad ascendente, y dentro de cada edad, ordenar por saldo descendente. Lo hacemos de la siguiente manera:

```
1 SELECT *
2 FROM bank_data
3 ORDER BY age ASC, balance DESC
4 ;
```

De esta manera, podemos observar que los clientes más jóvenes tienen en general menor saldo, por ejemplo:

	age	job	marital	educ...	default	balan...	hous...	loan	contact	day	month	durat...	camp...	pdays	previ...	pout...	y
3233	19	student	single	unkno...	no	1169	no	no	cellular	6	feb	463	18	-1	0	unkno...	no
2780	19	student	single	secon...	no	302	no	no	cellular	16	jul	205	1	-1	0	unkno...	yes
503	19	student	single	primary	no	103	no	no	cellular	10	jul	104	2	-1	0	unkno...	yes
1900	19	student	single	unkno...	no	0	no	no	cellular	11	feb	123	3	-1	0	unkno...	no
1725	20	student	single	secon...	no	1191	no	no	cellular	12	feb	274	1	-1	0	unkno...	no
13	20	student	single	secon...	no	502	no	no	cellular	30	apr	261	1	-1	0	unkno...	yes
999	20	student	single	secon...	no	291	no	no	teleph...	11	may	172	5	371	5	failure	no
4152	21	student	single	secon...	no	6844	no	no	cellular	14	aug	126	3	127	7	other	no
110	21	student	single	secon...	no	2488	no	no	cellular	30	jun	258	6	169	3	success	yes
2046	21	services	single	secon...	no	1903	yes	no	unkno...	29	may	107	2	-1	0	unkno...	no
2289	21	student	single	secon...	no	681	no	no	unkno...	20	aug	6	1	-1	0	unkno...	no
1391	21	services	single	secon...	no	361	no	no	teleph...	5	jun	329	1	95	1	other	no
2703	21	student	single	unkno...	no	137	yes	no	unkno...	12	may	198	3	-1	0	unkno...	no

Aunque también se puede ver que, entre los más mayores, hay clientes con saldo negativo, por ejemplo:

	age	job	marital	educ...	default	balan...	hous...	loan	contact	day	month	durat...	camp...	pdays	previ...	pout...	y
3740	36	blue-c...	divorced	primary	no	-308	yes	no	cellular	12	may	725	1	-1	0	unkno...	yes
514	36	mana...	married	tertiary	no	-381	yes	no	cellular	31	jul	30	29	-1	0	unkno...	no
1456	36	admin.	married	secon...	no	-423	yes	yes	unkno...	13	may	106	1	-1	0	unkno...	no
3610	36	techni...	divorced	secon...	no	-435	yes	no	unkno...	2	jun	85	8	-1	0	unkno...	no
857	36	blue-c...	married	secon...	no	-461	no	no	cellular	11	may	254	2	353	1	failure	no
3493	36	mana...	single	tertiary	no	-679	yes	yes	cellular	7	may	172	1	93	1	failure	no
4201	36	unem...	married	secon...	no	-872	yes	yes	cellular	20	nov	153	1	183	1	failure	no
2776	37	mana...	married	primary	no	22856	no	no	cellular	2	jul	154	1	388	1	failure	no
4334	37	entrep...	single	secon...	no	20453	yes	no	teleph...	4	may	115	1	-1	0	unkno...	no
875	37	admin.	married	secon...	no	11303	no	no	cellular	26	may	500	2	-1	0	unkno...	no
3921	37	entrep...	married	tertiary	no	7944	no	no	cellular	21	nov	102	1	-1	0	unkno...	no
4457	37	blue-c...	single	primary	no	6969	yes	no	unkno...	20	may	412	1	-1	0	unkno...	no
3192	37	techni...	married	tertiary	no	6968	no	no	cellular	1	jun	175	1	-1	0	unkno...	no

Otra forma más cómoda de ver esto sería seleccionando únicamente las columnas que nos interesan, en este caso la edad y el saldo, de la siguiente manera:

```
1 SELECT age, balance
2 FROM bank_data
3 ORDER BY age ASC, balance DESC
4 ;
```

Con lo cual nos sale una tabla más fácil de visualizar:

age	balance
19	1169
19	302
19	103
19	0
20	1191
20	502
20	291
21	6844
21	2488
21	1903
21	681
21	361
21	137

Vimos en nuestro anterior trabajo que la variable *poutcome* cuando tomaba el valor *success* era significativa para que el cliente se suscribiera. Vamos a probar a filtrar los datos por esta variable para ver si efectivamente hay mayor proporción de suscritos:

```
1 SELECT poutcome, y
2 FROM bank_data
3 WHERE poutcome = 'success'
4 ;
```

Tan solo echando un primer vistazo observamos que claramente hay una mayor proporción de clientes que se suscriben:

poutcome	y
success	yes
success	yes
success	no
success	yes
success	yes
success	no
success	yes
success	no
success	no
success	yes
success	yes
success	yes
success	no

Podemos compararlo con lo que sale si tomamos *poutcome* como *failure:*

```sql
SELECT poutcome, y
FROM bank_data
WHERE poutcome = 'failure'
;
```

Donde efectivamente comprobamos que hay una mayor proporción de clientes que no se suscriben:

poutcome	y
failure	no
failure	no
failure	no
failure	no
failure	no
failure	no
failure	no
failure	yes
failure	no
failure	no
failure	no
failure	yes
failure	no

También vimos en el anterior trabajo que los clientes contactados en marzo y octubre eran más propensos a suscribirse. Probemos a filtrar por resultado de la anterior campaña exitoso y por mes de marzo u octubre, y veamos si es cierto que se suscriben más clientes:

```sql
SELECT poutcome, month, y
FROM bank_data
WHERE poutcome = 'success'
AND (month = 'mar' OR month = 'oct')
;
```

Vemos que la mayoría de estos clientes sí se suscriben:

poutcome	month	y
success	oct	yes
success	oct	yes
success	mar	yes
success	mar	yes
success	oct	no
success	oct	yes
success	mar	yes
success	mar	yes
success	oct	yes
success	oct	yes
success	oct	no
success	mar	yes
success	oct	yes

Ahora pasamos a hacer algunas cuentas. Para empezar, vamos a contar el número de filas, es decir, de individuos, en nuestros datos:

```
1 SELECT COUNT(*)
2 FROM bank_data
3 ;
```

Obtenemos que hay 4521 individuos:

COUNT(*)
4521

A continuación, vamos a contar cuántos de estos clientes se suscriben y cuántos no. Para ello, podemos usar:

```
1 SELECT y, COUNT(*)
2 FROM bank_data
3 GROUP BY y
4 ;
```

De esta forma, obtenemos que se suscriben 521 clientes, mientras que 4000 no se suscriben:

y	COUNT(*)
no	4000
yes	521

Ahora vamos a probar a agrupar a los clientes por estado civil, y dependiendo de si se suscriben o no. Vamos a contar cuántos clientes hay en cada grupo, además de obtener el rango de edad en el que se encuentran y el saldo promedio en cada grupo. Para ello:

```
SELECT
    marital
    , y
    , COUNT(*) n_rows
    , min(age) age_min
    , max(age) age_max
    , avg(balance) balance_avg
FROM bank_data
GROUP BY marital, y
ORDER BY balance_avg DESC
;
```

En la siguiente tabla se observa que la mayoría de los clientes que se suscriben, 277, están casados, y que además los casados que se suscriben son los que tienen un saldo promedio mayor. De hecho, todos los que se suscriben, independientemente de su estado civil, tienen saldos promedio mayores. En cuanto a las edades, vemos que los casados y los divorciados en general son mayores que los solteros, aunque para un mismo estado civil no se ven grandes diferencias en los rangos de edad entre los clientes que se suscriben y los que no.

marital	y	n_rows	age_min	age_max	balance_avg
married	yes	277	24	87	1644.1624548736463
single	yes	167	19	69	1491.3353293413174
divorced	yes	77	27	84	1487.051948051948
single	no	1029	19	66	1455.396501457726
married	no	2520	23	86	1443.3035714285713
divorced	no	451	26	83	1060.1308203991132

Seguidamente, vamos a agrupar a los clientes que se suscriben por trabajo, además de observar sus saldos mínimos, máximos y promedio para cada uno de los trabajos de los clientes que se suscriben.

```sql
SELECT
    job
    , COUNT(*) n_subs
    , min(balance) balance_min
    , max(balance) balance_max
    , avg(balance) balance_avg
FROM bank_data
WHERE y = 'yes'
GROUP BY job
ORDER BY n_subs DESC
;
```

Los clientes que más se suscriben trabajan en la administración, seguidos por los técnicos, los obreros, los administrativos y los jubilados. Los desempleados son los menos suscritos.

job	n_subs	balance_min	balance_max	balance_avg
management	131	-970	12569	1733.5496183206108
technician	83	-824	11262	1425.7951807228915
blue-collar	69	-887	9328	1026.4927536231885
admin.	58	-306	6046	1331.6379310344828
retired	54	-1206	19317	2480.3333333333335
services	38	-477	7066	1112.657894736842
self-employed	20	6	6610	1546.7
student	19	0	5291	1198.2631578947369
entrepreneur	15	-701	3904	943.1333333333333
housemaid	14	0	26965	3900.3571428571427
unemployed	13	94	3391	1336.076923076923
unknown	7	0	4717	1349.857142857143