Data Cleaning:

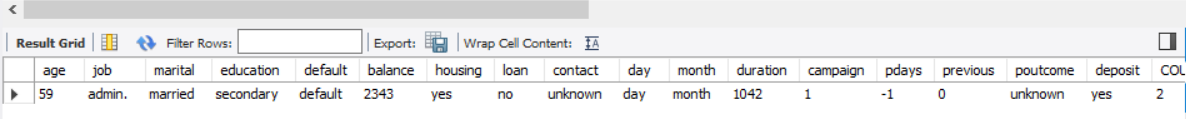
***Finding duplicates:***

SELECT age, job, marital, education, 'default', balance, housing, loan, contact, 'day', 'month', duration, campaign, pdays, previous, poutcome, deposit, COUNT(\*)

FROM BANK\_marketing

GROUP BY age, job, marital, education, 'default', balance, housing, loan, contact, 'day', 'month', duration, campaign, pdays, previous, poutcome, deposit

HAVING COUNT(\*) > 1;



We only have 1 result.

***Identifying ‘id’ of the duplicate***

WITH RankedDuplicates AS (

SELECT id, age, job, marital, education, `default`, balance, housing, loan, contact, `day`, `month`, duration, campaign, pdays, previous, poutcome, deposit,

ROW\_NUMBER() OVER (PARTITION BY age, job, marital, education, `default`, balance, housing, loan, contact, `day`, `month`, duration, campaign, pdays, previous, poutcome, deposit ORDER BY id) AS row\_num

FROM

BANK\_marketing

)

SELECT id

FROM RankedDuplicates

WHERE row\_num > 1;

Id = 2.

***Deleting the duplicate***

* Add Primary Key Constraint to avoid the safe update mode

ALTER TABLE BANK\_marketing

ADD PRIMARY KEY (id);

* Delete the duplicated record

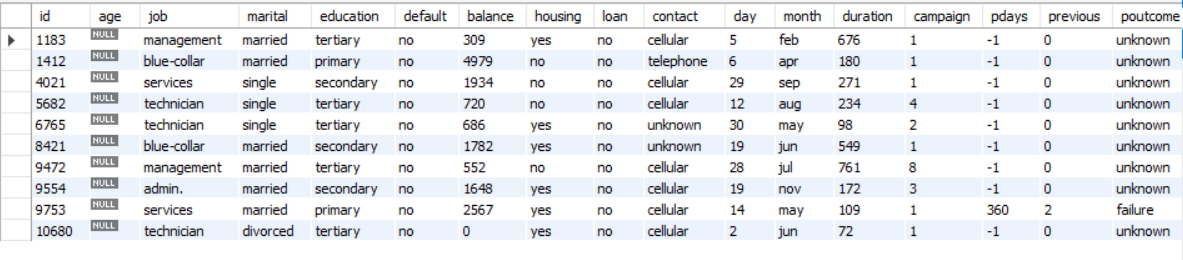
DELETE FROM BANK\_marketing

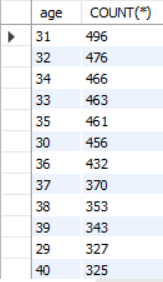
WHERE id = 2;

***Handling missing or blank values:***

*By age column:*

*We found 10 customers with unknown age.*





*Should we “populate” this info? And how?*

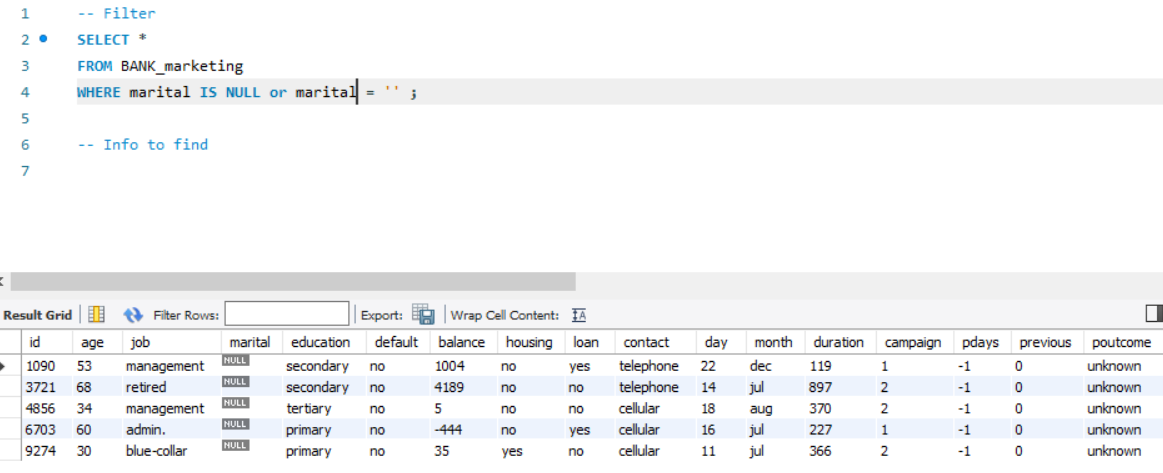
*I think the solution that fits better is using the median grouping by job and marital status.*

*I think using the median just for everybody wont be that much accurated, so I prefer to do a median about the position and if he/she is single or not. Single people for example are more younger, than divorced people.*

*By marital status:*

* type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

*5 rows returned from almost 12k rows*

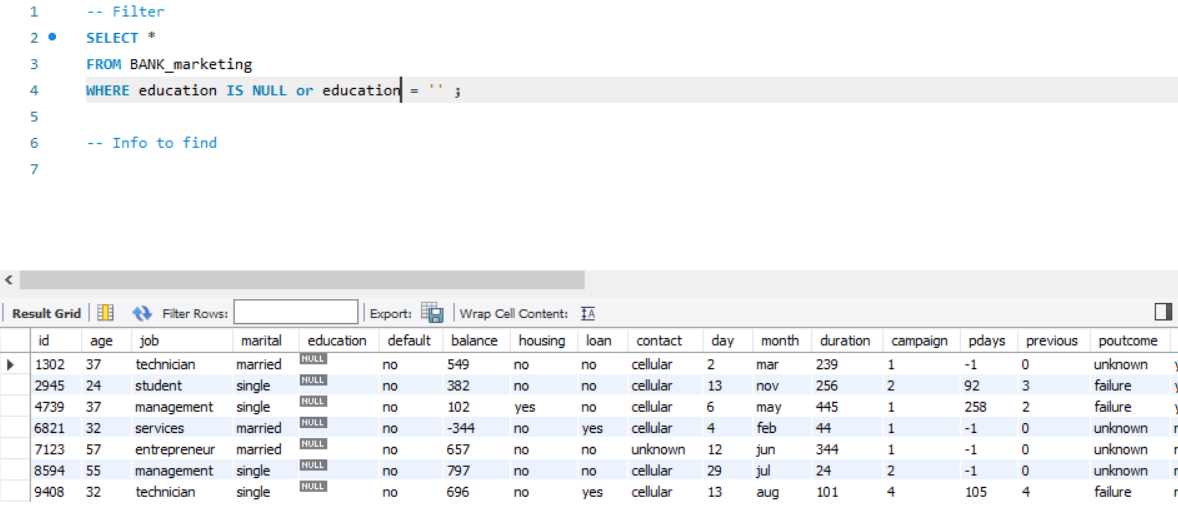


*First impression is, trying to solve this but require more time than expected and also there are only 5 persons with no marital status, so we are not losing so much info here.*

*By education:*

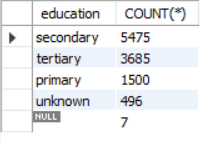
* (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

*7 rows returned*



*For only 7 customers would be imposible to know for different reasons.*

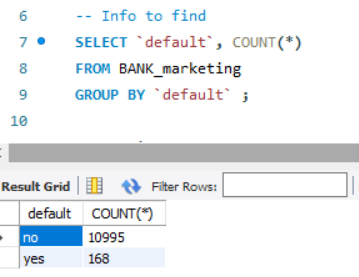
1. *We have also another problem that is the we have the value “unknown”. We have 496 employees that we don’t know where they are working.*
2. *Also, when you check the difference educations we also have find out that are very distributed.*



*By default column:*

* *The default column means if the employee or customer is in suspension of his/her legal duties, So this person still needs to pay some bills.*

*Doesn’t appear any NULL or blank Value.*



*By Balance column:*

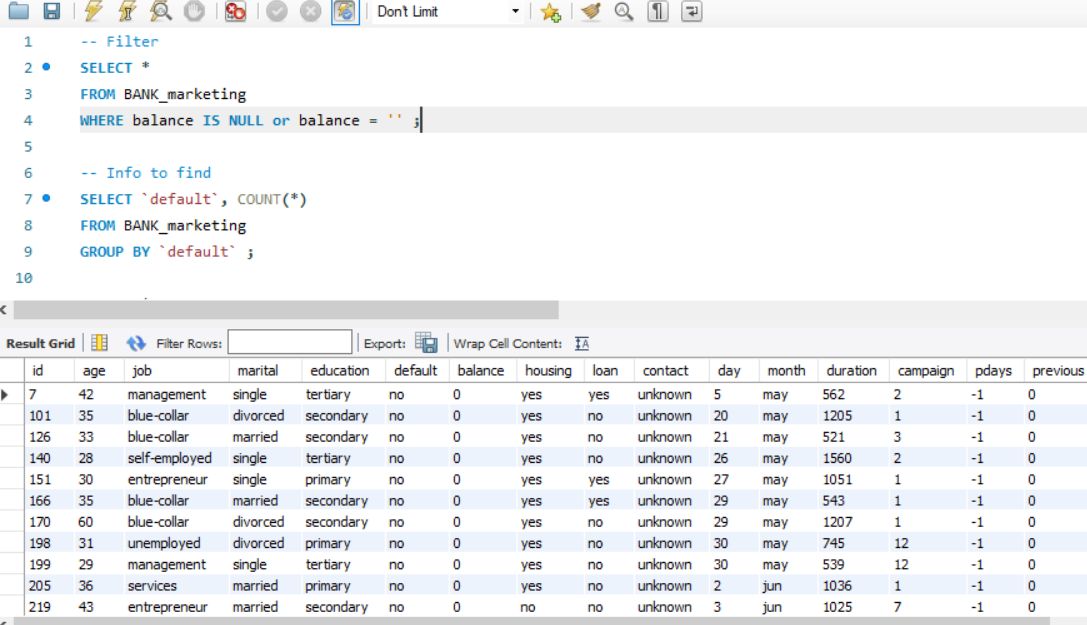
* average yearly balance

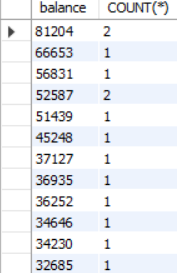
*774 rows returned*

*The thing about the zero values could mean 2 things:*

*More probably: There was not an input on this row, so integer format put a zero automatically so we having a mix of people that we don’t know their balance and also a few people who casually their balance is real zero.*

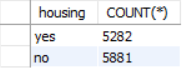
*Less probably: I find it hard but could happens that there are 774 employees with 0 real balance.*





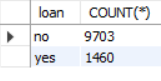
*By Housing column: “Mortgage”*

*Doesn’t appear any NULL or blank Value.*



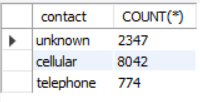
*By Loan Column: “Prestamo”*

*Doesn’t appear any NULL or blank Value.*



*By Contact Column:*

*Doesn’t appear any NULL or blank Value.*



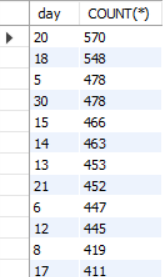
*Questions:*

*Can be populated the unknow value?*

*Does cellular and telephone is the same?*

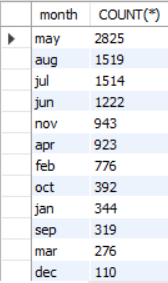
*By Day Column:*

*Doesn’t appear any NULL or blank Value.*



*By Month Column:*

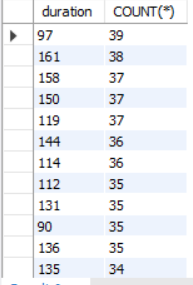
*Doesn’t appear any NULL or blank Value.*



*By Duration Column:*

|  |  |
| --- | --- |
|  | last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. |

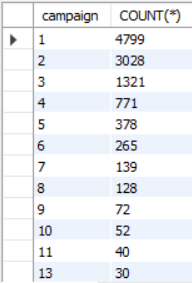
*Doesn’t appear any NULL or blank Value.*



*By Campaign Column:*

* number of contacts performed during this campaign and for this client (numeric, includes last contact)

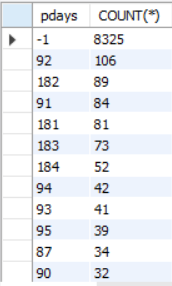
*Doesn’t appear any NULL or blank Value.*



*By Pdays Column:*

* number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)

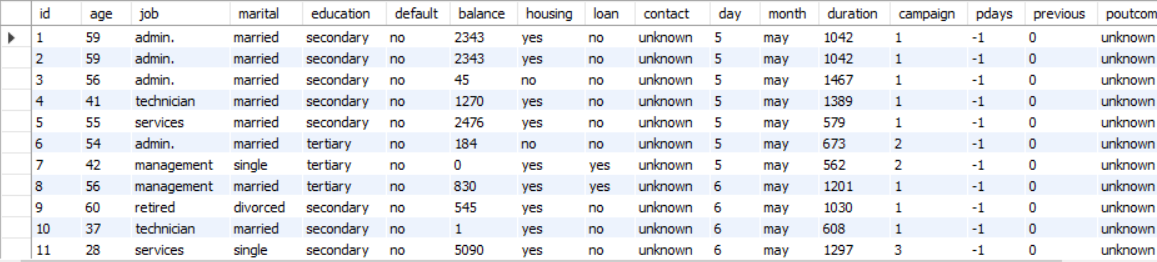
*Doesn’t appear any NULL or blank Value.*

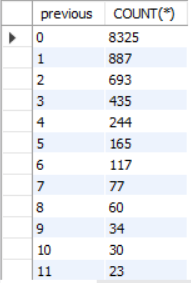


*By Previous Column:*

* number of contacts performed before this campaign and for this client

*We found more than 8k rows with 0 value.*





*Questions:*

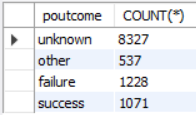
*Zero means that in the database there was no input?*

*Or means that the employee didn’t receive any call?*

*By Poutcome Column:*

* outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

*Doesn’t appear any NULL or blank Value.*

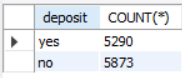


*Questions:*

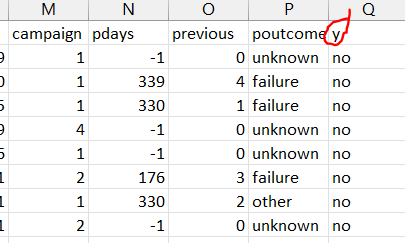
*Is this any way we can replace the unknow value?*

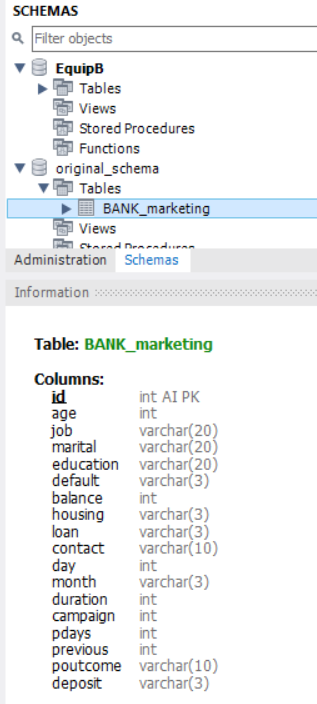
*By Deposit Column:*

*Doesn’t appear any NULL or blank Value.*



* *Seems like we lost a column in the external database:*

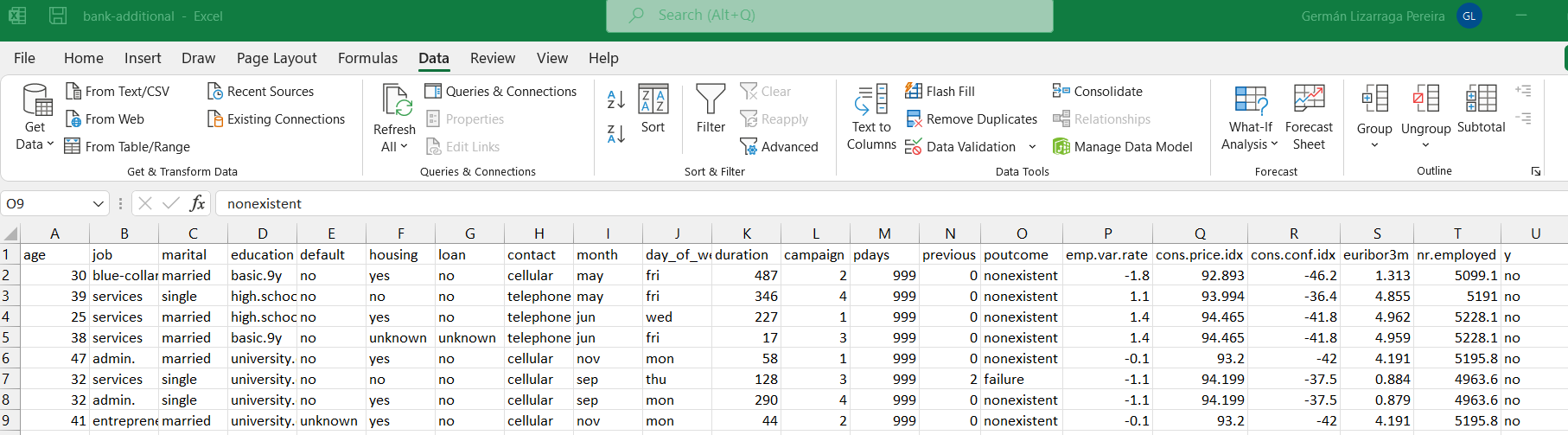




*If the excel and the external database has the exact same rows I guess we can insert this without any problem. In case we insert this info is important to know that the source file is bank-full.csv.*

*Just reminder that the file bank.csv is just 10% info from bank-full.csv*

*Also there is an additional excel*



*I’m not quite sure if we will have to merge this info in the original table or create a new one just with the info needed.*