Data Cleaning - Bank Marketing Dataset

#### **1. Introducción**

En este documento se detalla el proceso de limpieza de datos llevado a cabo sobre el dataset. La limpieza de datos es una etapa fundamental para asegurar la integridad, precisión y consistencia de los análisis posteriores.

#### **2. Objetivo de la Limpieza de Datos**

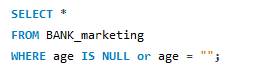
El objetivo del proceso de limpieza de datos es garantizar que los datos sean de alta calidad y estén listos para el análisis. Este proceso implica la eliminación o corrección de valores erróneos, el tratamiento de datos faltantes, la estandarización de las variables y la detección de posibles anomalías o outliers.

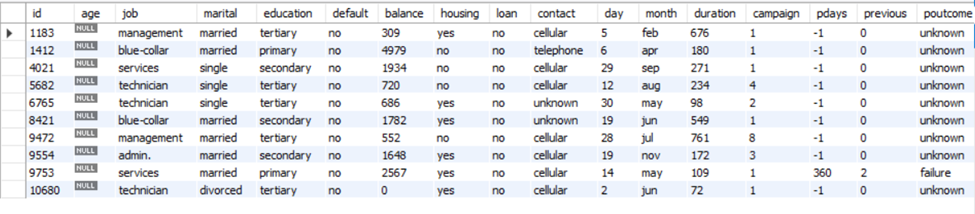
#### **3. Pasos del Proceso de Limpieza**

##### **3.1. Identificación y Tratamiento de Valores Nulos**

Se inspeccionaron las diferentes columnas del DataSet para identificar los valores nulos de cada una.

**Identificación de valores Nulos por “Age”**

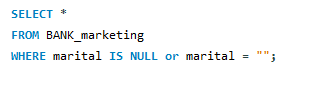
****

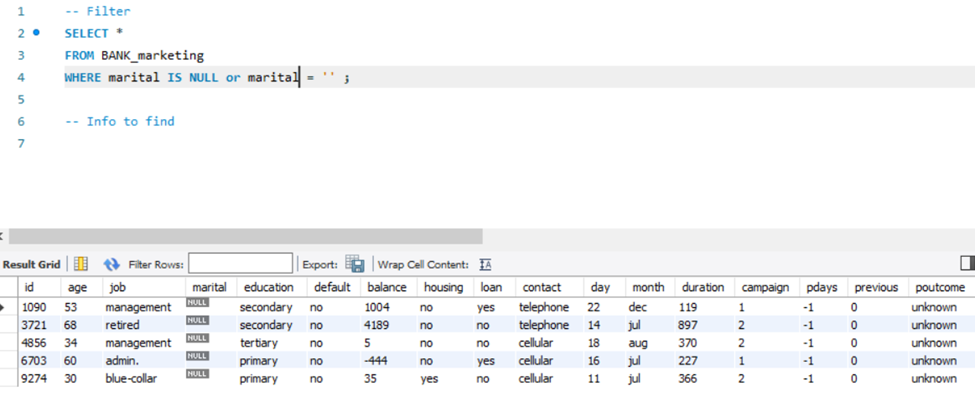
****

Se obtuvieron **10 clientes** con edad desconocida (**unknown**)

* **Acción tomada**:
  + Se utilizó la mediana de edad agrupada por trabajo(job) y estado civil (marital status)

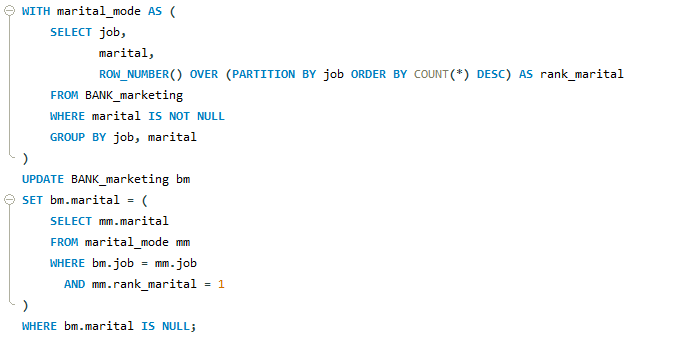
**Identificación de valores Nulos por “marital status”**

****

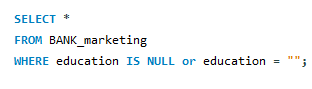
****

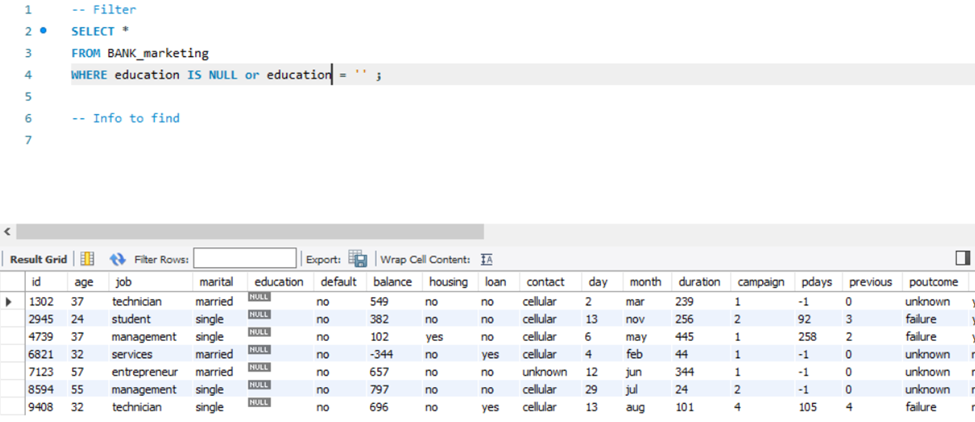
Se obtuvieron **5 clientes** con **valor nulo** para la columna marital.

* **Acción tomada**:
  + Se utilizó la moda de edad para cada tipo de trabajo

****

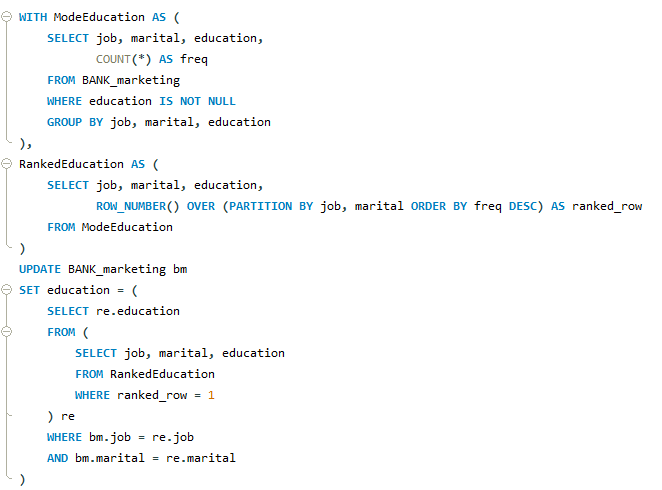
**Identificación de valores Nulos por “Education”**

****

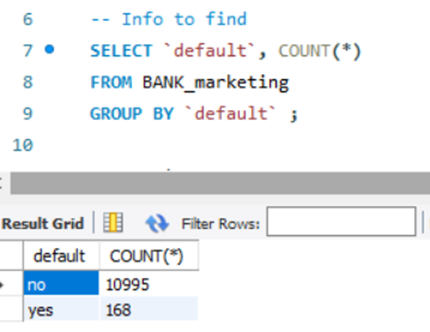
****

Se obtuvieron **7 clientes** de los cuales no tenemos información de su educación.

* **Acción tomada**:
  + Se utilizó la moda de educación agrupada por estado civil y trabajo.

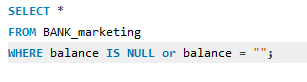


**Identificación de valores Nulos por “default”**

****

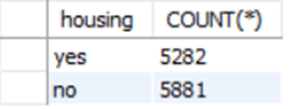
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Balance”**

****

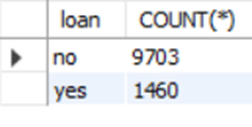
Se obtuvieron 774 registros en los que el “balance” es 0.

**Identificación de valores Nulos por “Housing”**

****

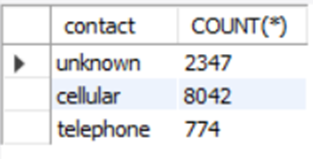
**No se obtuvieron valores nulos**

**Identificación de valores Nulos por “Loan”**

****

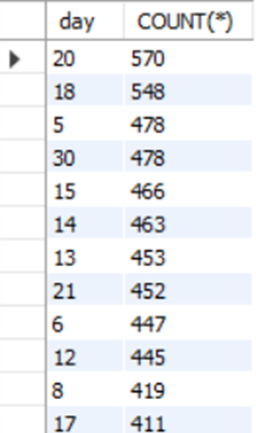
**No se obtuvieron valores nulos**

**Identificación de valores Nulos por “Contact”**

****

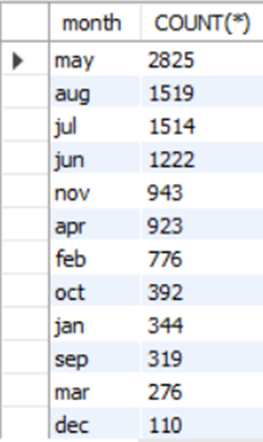
**No se obtuvieron valores nulos.** Pero tenemos **2347 “unknowns”.**

**Identificación de valores Nulos por “Day”**

****

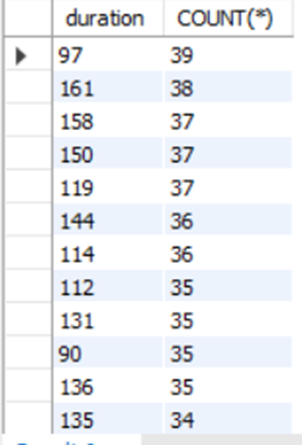
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Month”**

****

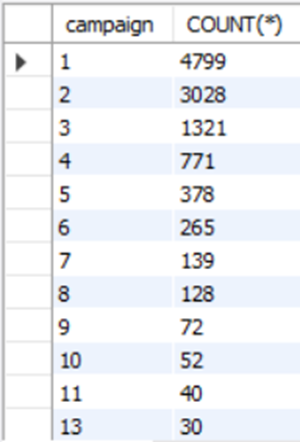
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Duration”**

****

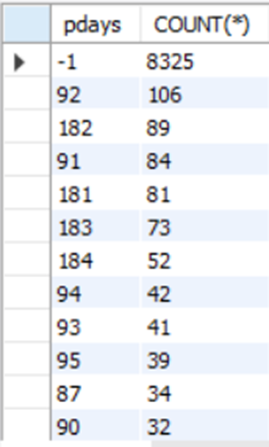
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Campaing”**

****

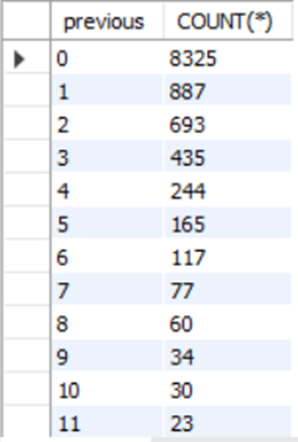
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Pdays”**

****

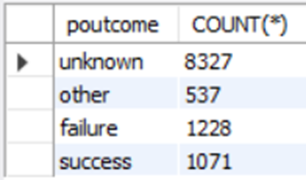
**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Previous”**

****

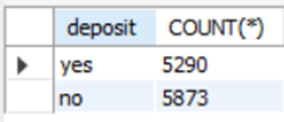
**No se obtuvieron valores nulos**, pero tenemos más de 8000 registros con valor 0. ¿Significa esto que no recibieron ninguna llamada? ¿O es un problema de la base de datos?

**Identificación de valores Nulos por “Poutcome”**

****

**No se obtuvieron valores nulos.**

**Identificación de valores Nulos por “Deposit”**

****

**No se obtuvieron valores nulos.**

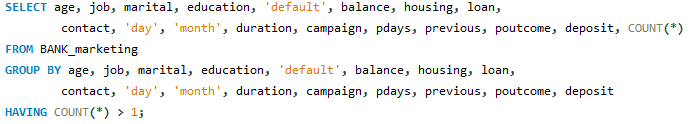
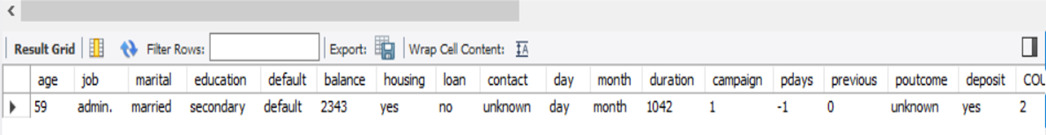
##### **3.3. Eliminación de Datos Duplicados**

Se revisó el dataset para detectar y eliminar filas duplicadas que pudieran distorsionar los resultados.

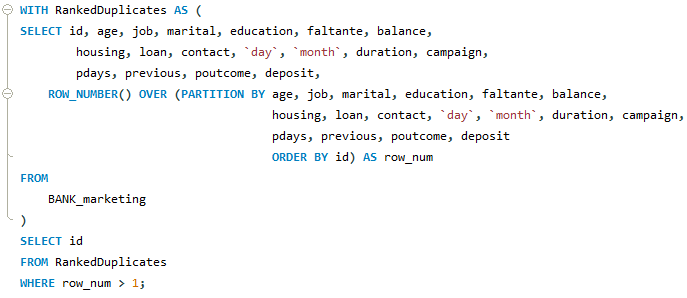
* **Acción tomada**:

Se ejecutó un proceso de detección de duplicados basado en múltiples columnas clave. Los duplicados fueron eliminados para evitar la repetición de clientes en el análisis.

**Detección de duplicados:**

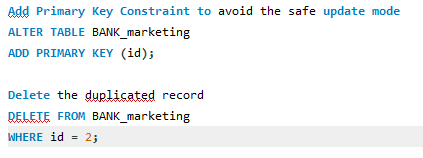
  


**Identificación del ID duplicado:**



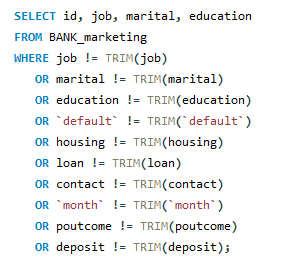
**ID= 2**

**Eliminación del duplicado:**

****

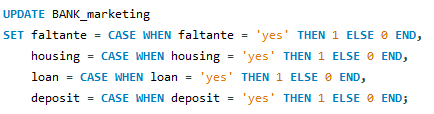
**3.4. Revisar espacios**

Se revisó que en ninguna columna aparecieran espacios en blanco ni por delante ni por detrás.

****

**3.5 Redefinir el tipo de dato**

Se substituyó el “yes” y el “no” por 1 y ceros en las variables “faltante”, “housing”, “loan” y “deposit”.



A continuación, se cambió el tipo de dato de estas columnas a booleano.

