# Data Cleaning and Transformation Notes

In this phase we have performed **Data Cleaning and Transformation**: managing duplicates, irrelevant columns or rows, incorrect data typing, data normalization (same language, upper-/lower-casing, etc.).

1. First of all, we have **removed irrelevant rows**, meaning those which contain duplicated information or no information at all. For this purpose, we use have used the bottom Remove Duplicates. 14 duplicates where removed. No rows without any information were observed.
2. Then, we jumped into **irrelevant columns**, those which only generate noise and were not going to be used. We checked whether the column Churned was the same as Churned Label. For that reason, we created an extra column to indicate OK or Not OK if the following conditions were met:

=SI(O(Y([@[Churn Label]]="No";[@Churned]=0);Y([@[Churn Label]]="Yes";[@Churned]=1));"OK";"Not OK")

Thanks to the filter of that extra column we observed that no rows were tagged as “Not OK” so one of the columns could be removed as they stood for the same. We decided to remove the field Churned because we wanted to maintain the tags (Yes/No) for the final dashboard. For the transformation phase, we thought it would be interesting to convert the additional Boolean fields in Yes/No tags, if applicable.

We have also removed the column Phone number, as it was irrelevant for our analysis.

We observed that there are some inconsistencies regarding the different categories of the field Gender. Two of them are in English (Female / Prefer not to say) and the remaining is in Spanish (Hombre).

Then, we decided to maintain the columns Under30 and Senior because we thought they might be related with a KPI of our client (the alternative was to remove both and construct our own age groups from the field Age). However, we first wanted to validate those fields by comparing their (Yes/No) tags with the actual value in the field Age. For that reason, we created extra columns to indicate OK or Not OK if the following conditions were met:

=SI(O(Y([@Age]>=30;[@[Under 30]]="No");Y([@Age]<30;[@[Under 30]]="Yes"));"OK";"Not OK")

=SI(O(Y([@Age]>65;[@Senior]="Yes");Y([@Age]<=65;[@Senior]="No"));"OK";"Not OK")

Tags values of the field Under30 were consistent with the age values of the field Age. However, some values of the field Senior were inconsistent. In some cases, 65 years old subscribers were tagged as Senior (Yes) and others were tagged as not Senior (No). We decided not to modify that column before asking the client to clarify those inconsistencies.

Taking advantage of those formulas, we also performed data validation to the field Number of Customers in Group, to check whether its values were consistent with the Yes/No tags of the field Group:

=SI(O(Y([@Group]="No";[@[Number of Customers in Group]]<2);Y([@Group]="Yes";[@[Number of Customers in Group]]>1));"OK";"Not OK")

All the values were correct.

After asking the client about the field Customer Tenure (in months), they answered that field was not relevant for the analysis as it had nothing to do with the subject. Therefore, we decided to remove it.

Other irrelevant columns that were removed: Hobby, Favorite TV Show, Interna Notes.

1. The next step was to **manage incorrect data typing**. The field Monthly Charge contained some values with the USD currency suffix and others without the suffix. We applied the following nested formulas in a new column (Monthly Charge (USD)) to extract only the numbers and turn them into number data type:

=VALOR.NUMERO(EXTRAE([@[Monthly Charge]];1;SI.ERROR(HALLAR(" ";[@[Monthly Charge]]);[@[Monthly Charge]])))

The original Monthly Charge field was hidden.

Next, we modified the field name from Contact Date to Contact Date Hidden and created a new column called Contact Date which turned the values of the field Contact Date Hidden into date type using the following formula:

=FECHANUMERO([@[Contact Date Hidden]])

Later, we hid the field Contact Date Hidden.

To end with the correction of data types, we transformed the Avg Monthly Expenses field values into number type and reduced the number of decimals into two.

1. Once data types were correctly established, we tried to perform **data normalization**. Starting with the Unlimited Data Plan field, we decided to modify the field name into Unlimited Data Plan Hidden and created a new column called Unlimited Data Plan which turned the Boolean values into Yes/No tags, which are more suitable for the final dashboard:

=SI([@[Unlimited Data Plan Hidden]]=0;"No";"Yes")

Then, we hid the field Unlimited Data Plan Hidden.

Moreover, we normalized the categories of the field Gender ensuring that all of them were written in English. For that purpose, we changed the field name to Gender Hidden and created a new column called Gender, where we applied the following conditional formula:

=SI([@[Gender Hidden]]="Hombre";"Male";[@[Gender Hidden]])

Then, we hid the field Gender Hidden.

Next, on the field Payment Method, we decided to capitalize the first letter of each word as if the values were proper nouns. For this aim, we changed the field name to Payment Method Hidden and created a new column called Payment Method. We used the following function:

=NOMPROPIO([@[Payment Method Hidden]])

Finally, for the fields Customer Segment, Preferred Contact Method and Applied Discount, we also wanted to ensure that all value categories were written in English. In order to achieve that goal, we created an auxiliary table at a different sheet. The table consisted on two columns, one for the term in Spanish (Origen) and another for the same term in English (Resultado):

Tabla3:

|  |  |
| --- | --- |
| **Origen** | **Resultado** |
| Alto | High |
| Medio | Medium |
| Bajo | Low |
| Teléfono | Phone |
| VERDADERO | Yes |
| FALSO | No |

Thanks to the following formulas, we managed to substitute Spanish terms into English ones:

=BUSCARV([@[Customer Segment Hidden]];Tabla3[#Todo];2;FALSO)

=SI.ERROR(BUSCARV([@[Preferred Contact Method Hidden]];Tabla3[#Todo];2;FALSO);[@[Preferred Contact Method Hidden]])

=BUSCARV([@[Applied Discount Hidden]];Tabla3[#Todo];2;FALSO)

Finally, we checked again for duplicated rows. It is common that, after performing data cleaning and transformations, additional duplicated rows appear. In fact, 156 additional duplicated rows were identified and removed from the dataset.