Data Cleaning

El proceso de limpieza de datos tiene una serie de pasos a seguir para poder realizarlo de forma eficiente y ordenada (aunque puede hacerse en casi cualquier orden).

Todos los notebooks tienen una sección denominada “Concatenación con el .csv enviado” que refiere a justamente levantar el .csv que se guardó previo al envío de datos al proveedor (aquel que contiene IDMOROSO, IDCLIENTE, ESTADO, FILTROA y DATOADIC si corresponde).

En cada notebook se tendrán que realizar 3 modificaciones generales:

1. Modificar la ruta del .csv/excel con los datos recibidos por el proveedor, al inicio del notebook.
2. Modificar la ruta correspondiente en la sección “Concatenación con el .csv enviado”, apuntando al archivo generado para el envío (aquel que contiene IDMOROSO, IDCLIENTE, ESTADO, FILTROA y DATOADIC si corresponde).
3. Modificar la ruta correspondiente al export final, al final de cada notebook.

Se tendrán que realizar ciertas modificaciones adicionales si:

- El proveedor agrega o quita columnas del dataset. Esto va a repercutir en ciertos índices en algunas columnas, que habrá que corregir.
- El proveedor renombra ciertas columnas.
- El proveedor cambia el tipo de dato que envía por columna; por ejemplo, en una determinada columna puede enviar floats, pero al envío siguiente puede enviar integers.
- Se incluye a AMEX en el envío; se tendrá que tomar el DNI de la columna DATOADIC en lugar de la columna IDMOROSO. Estas funciones están disponibles en los notebooks, simplemente hay que descomentarlas según el caso.

Limpieza de Datos de Contactabilidad

En primer lugar, se ejecuta el notebook que refiere a **Data Cleaning - Mails**. Los mismos van a correr ciertos procesos que finalizan con un export de un archivo .csv, con las columnas ya adaptadas a la tabla en la que se insertará, denominada DAM_CONTACTABILIDAD.

En segundo lugar, se ejecuta el notebook referido a **Data Cleaning - Teléfonos**. Este notebook contiene algoritmos de limpieza de datos para adecuar los datos telefónicos a la tabla DAM_CONTACTABILIDAD.

Duplicados

En algunos casos, algunos proveedores nos devuelven duplicados ciertos teléfonos o mails de la siguiente forma:

- El teléfono 1111-1111 aparece en el campo “Teléfono Fijo Personal” pero también en el campo “Teléfono Fijo Laboral”. Por lo tanto, no tenemos un propietario al que podamos atribuirle con certeza ese dato.
- Es por esto, que si al final de los pipelines Data Cleaning - Teléfonos o Data Cleaning - Mails encontramos estos duplicados, los mismos deberán pasar a un pipeline llamado **DAM_CONTACTABILIDAD - Teléfonos - Duplicados**.
- En ese pipeline, simplemente se generan puntajes en función de ciertas reglas definidas, para elegir a quién finalmente atribuimos ese dato.
- El output de este pipeline es el que finalmente se inserta en la tabla DAM_CONTACTABILIDAD.

Limpieza de Datos Patrimoniales

Para aquellos proveedores que nos brindan datos patrimoniales (PyP/VIS e Insiders), se ejecutan los siguientes notebooks, preferentemente en el siguiente orden:

Los siguientes notebooks se ejecutan utilizando el archivo original que recibimos por parte del proveedor.

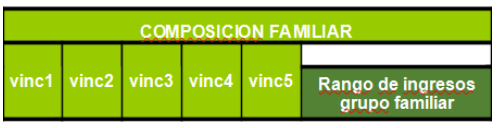
- Data Cleaning - Inmuebles
- Data Cleaning - Vehículos

IMPORTANTE: Para los casos que aparecen a continuación, se deberá generar una copia del archivo que nos devuelve el proveedor, y trabajaremos sobre esa copia.

Debido a que tenemos muchas celdas mergeadas en el header del archivo, la idea es desmergearlas para poder identificar más rápidamente cada campo (a futuro, es un proceso completamente automatizable).

Además, dentro del archivo recibido, tenemos columnas que tienen el mismo nombre pero la información que nos otorgan es diferente: por ejemplo, tenemos dos celdas que se llaman Cant. Sit. 1, de las cuales una de ellas pertenece al titular, y otra al grupo familiar. Si bien ambas están enmarcadas dentro de un título de mayor jerarquía, para preservar la integridad del proceso se decide renombrarlas. Entonces:

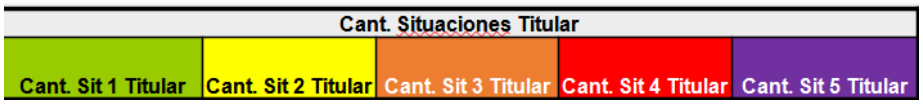
1. Se genera una copia del archivo que nos devuelve el proveedor.
2. Se desmergean las celdas y se mueven a la celda inferior como se observa en la imagen:
    a. Rango de Ingresos (dentro de Composición Familiar).
    b. Desde Propiedades hasta Patentados menor al año 2001 (pasando por todos los patentados de todos los años, dentro del título Bienes Registrados (Titular)).
    c. Rango de Ingresos (dentro de AFIP - Autónomos - Monotributistas).
    d. Rango de Ingresos (dentro de Haberes).



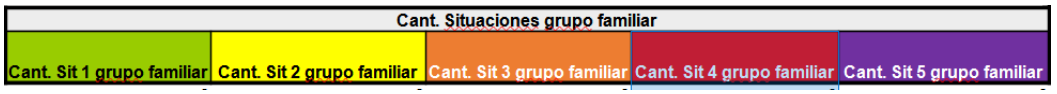
e. Ingresos Mes 6 (dentro de Rango de Ingresos últimos 6 meses).

3. Se renombran algunas celdas:

   a. **Situaciones**: Dentro de Comportamiento en Banco Central Titular, se renombra Cant. Situaciones a Cant. Situaciones Titular. Se deberá realizar lo mismo con cada situación que aparece posteriormente en ese grupo.



   b. Deberá realizarse lo mismo con la Cantidad de Situaciones para el grupo familiar.



   c. Dentro de Composición Familiar >> Rango de Ingresos, se renombra “Rango de Ingresos” a “Rango de ingresos grupo familiar”.
   d. Dentro de AFIP-Autónomos-Monotributistas, se renombra “Rango Ingresos” como “Rango Ingresos autonomo”.
   e. Dentro de Haberes, se renombra “Rango de Ingresos” como “JUBILACION - PENSION”.
   f. Dentro de Principales Entidades con las que Opera el Titular se renombran “Entidad n” a “Entidad n Titular”, y “Deuda n” a “Deuda n Titular”, con n=1,2,3,4.

Si algún paso se omite, un error dentro del pipeline indicará qué columna está generando problemas (porque no va a poder encontrarla). Todos estos cambios de nombre para las columnas, también podrán ser observados dentro de los notebooks, ya que los mismos van a buscar las columnas ya renombradas.

Una vez realizado esto (es algo que con algo de tiempo se puede automatizar a futuro), se continúa ejecutando los siguientes notebooks:

- Data Cleaning - Deudas (*)
- Data Cleaning - Scoring

(*) Deprecated para el proveedor Insiders. No tenemos una definición certera sobre qué situación corresponde a qué entidad, por lo que por el momento no se utiliza.

Finalmente, se ejecuta el último pipeline:

- Data Cleaning - DAM_LOG

Este pipeline simplemente cruza los datos enviados con los recibidos, para obtener aquellos casos que fueron enriquecidos, y generar así un log. Notar que el primer dataset que se procesa es el de casos enviados, y luego el de casos recibidos.