# **Proceso de adaptación de BD para población total**

El presente resumen tiene como objetivo complementar el proceso de tratamiento y ajuste de los datos de los archivos correspondientes al ETL llamado “*Población\_H\_M\_T\_2013.xlsx” y* “*Población\_H\_M\_T\_2017.xlsx”*, a utilizar por el grupo “Politics”, además de entregar una mejor comprensión sobre el código utilizado para este fin.

Dado que el análisis exploratorio de datos se centró principalmente en la obtención de información relevante para el estudio, este solo fue posible realizarlo a partir de la única variable cuantitativa con la cual contaba el *dataset,* correspondiente a la *cantidad de votos obtenidos por cada candidato.* Dado lo anterior, y con la finalidad de realizar comparaciones entre las diferentes comunas y/o regiones fue necesario realizar una segunda versión del *dataset*, que en esta instancia contendría información relevante correspondiente al total de población por cada comuna y región.

Si bien para ambos procesos de elecciones el tratamiento realizado debió ser diferente, a pesar de utilizar información estadística de una fuente confiable como lo es INE, las bases de datos no eran totalmente coincidentes, tanto en su formato como en sus resultados. Esto, principalmente porque para el año 2013 se debió utilizar una proyección, mientras que para el año 2017 se utilizaron estadísticos emanados desde el censo como se muestra a continuación.

**ETL V2 Elecciones 2017:**

En este caso, se hizo uso de la base de datos oficial publicada por el INE correspondiente a la entrega final de los resultados definitivos del Censo del año 2017. Esta BD contenía los valores correspondientes al total de población clasificados por: **comunas, provincias y regiones.**

Para el caso de nuestro análisis en particular solo se utilizó la clasificación correspondiente a **comunas**, la cual además abarcaba los atributos manejados las otras clasificaciones, estos eran los siguientes:

* **Nombre Región:** Atributo String.
* **Código región:** atributo numérico.
* **Nombre Provincia:** Atributo String.
* **Código provincia:** atributo numérico.
* **Nombre comuna:** Atributo String.
* **Código comuna:** atributo numérico.
* **Edad:** atributo numérico, cada fila corresponde a un intervalo que agrupa 5 años.
* **Hombre:** atributo numérico.
* **Mujeres:** Atributo numérico.
* **Total:** suma de ambas columnas anteriores.

De los atributos anteriores, solamente se seleccionaron aquellos relevantes para el análisis eliminando todos los demás, finalmente los atributos utilizados correspondieron a: **Nombre comuna** (el cual posteriormente fue nombrado **comuna\_name**), **edad, hombres, mujeres, total.**

El proceso anterior generó la base de datos, *Población\_H\_M\_\_T\_2017*.*xlsx*, preparada para ser utilizada en conjunto con las bases de datos correspondientes a la V1 del ETL, como se presenta a continuación:



**Figura 1-** Base de datos poblacional elecciones 2017.

La base de datos original puede ser consultada [aquí](http://www.censo2017.cl/wp-content/uploads/2017/12/Cantidad-de-Personas-por-Sexo-y-Edad.xlsx).

**ETL V2 Elecciones 2013:**

En el caso del proceso de elección correspondiente al año 2013, se hizo uso de la base de datos de estadísticas de proyección demográficas y vitales correspondientes al año 2013, la cual proyecta el total de población por comuna, su correspondiente región, y entregando proyecciones totales para cada año de edad, desde los 0 hasta los 80 años, los atributos que maneja esta BD son los siguientes:

* **Región:** atributo numérico.
* **Nombre\_region:** atributo String (considerando caracteres especiales).
* **Provincia:** atributo numérico.
* **Nombre\_provincia:** atributo String (considera caracteres especiales).
* **Comuna:** Atributo numérico, correspondiente a un identificador.
* **Nombre\_comuna:** atributo String.
* **Sexo:** Atributo numérico binario (1= hombres, 2=mujeres).
* **Edad:** atributo numérico, cada fila corresponde a un año diferente.
* **Proyecciones:** con el formato aXXXX donde el valor posterior a la letra inicial corresponde al año de la proyección.

Dada la gran cantidad de atributos con los que cuenta la base de datos inicial, y aquellos que son de interés para nuestro análisis en primera instancia se eliminó todas aquellas columnas no relevantes, para ello solo se mantuvieron: **nombre\_comuna, Sexo (binario), edad, a2013 (correspondiente al periodo relevante).**

Por simplicidad en el manejo de los datos y con la finalidad de disminuir el número de filas y dado que la proyección estaba dividida por cada año y sexo (cada comuna contaba con 162 filas correspondiente a población), se agruparon las proyecciones en intervalos de edad cada 5 años. Esto también con la finalidad de estandarizar el análisis para ambos procesos eleccionarios, finalmente, se establecieron los intervalos como se presentan a continuación:



**Figura 2-** Estandarización intervalos de edad elecciones 2013.

A través de un tratamiento mediante Excel y por medio de tablas dinámicas se realizó la clasificación de la población por intervalos de edad, separado por sexo y además se obtuvo el total de cada uno de los intervalos como se presenta en la siguiente figura, para cada comuna:



**Figura 3-** Estandarización intervalos de edad elecciones 2013.

El proceso anterior generó la base de datos, *Población\_H\_M\_\_T\_2013*.*xlsx*, preparada para ser utilizada en conjunto con las bases de datos correspondientes a la V1 del ETL, como se presenta a continuación:

La base de datos original puede ser consultada [aquí](http://www.ine.cl/docs/default-source/demogr%C3%A1ficas-y-vitales/demograf%C3%ADa/proyeccion-vigente/proyecciones/base_2002a2020_v3.xls?sfvrsn=10).