Se trata de un gimnasio. Se quiere obtener una caracterización de los clientes actuales, en función de la edad, sexo, tipo de membresía, código postal.

Para ello, se han recibido un ficheros en formato csv con datos sobre las membresías.

El gimnasio tiene datos desde el mes de julio de 2024 que fue cuando se migró al nuevo CRM.

Análisis EDA

**1. Carga de datos**

Mediante PowerQuery->Obtener datos

En la Hoja de Activos 06\_12, fecha en la que se realiza la descarga de datos.

**2.Identificación de variables**

De todas las variables, se describen las variables que se van utilizar:

* Member ID: número entero, identificor único de cada Cliente en la plataforma
* Gender: sexo, M, F, U (Undefined, no definido). El valor U no debería existir y se ha indicado al gimnasio que deberían corregirlo en M o F.
* Birthday: fecha de nacimiento, formato fecha
* ZIP Code: código postal
* City: texto, Ciudad
* Subscription reason: razón por la que el cliente se suscribió, variable categórica
* Source: cómo conoció al gimnasio
* Memberships: tipo de producto contratacto, categórica
* Member since: fecha de alta
* Contract end date: fecha de fin, al ser un reporte con activos y membresía mensual, todos tienen el final de mes
* Price: cuota, entero, número de cuatro cifras sin decimales (habrá que transformar)

**3. Transformación y limpieza de datos**

**3.1 Identificación de duplicados**

.-Sobre la columna Member ID, primero se aplica formato condicional para marcar en rojo si hay valores duplicados, ninguno está en rojo

.-También se realiza la función "CONTAR.SI" para contar las veces que aparece un identificador, restándole la primera ocurrencia, y es cero

.-En el menún tabla se clicka en "Eliminar duplicados" y devuelve el mensaje "No se encontraron duplicados"

**3.2 Valores faltantes/nulos**

Con la función Excel CONTARBLANCO se cuentan los campos que están en blanco en cada columna.

Acciones

"Suscription Reason": se duplica la columna por otra columna "Suscription Reason C", y los datos faltantes se sustituyen por "Desconocido"

"Source": se duplica la columna por otra que pone "Source C", y los datos faltantes se sustituyen por "Desconocido"

"Contract End Date": se duplica la columna por otra que pone "Contract End Date C", y los datos faltantes se sustituyen por la fecha del final de mes actual, dado que todas las membresías son mensuales autorenovables

**3.3 Transformación de datos**

"Price", esta variable es un número entero de 4 cifras, se crea una columna "Price C", que convierte el valor de la columna "Price" a euros, es la cuota mensual con iva

**3.4 Variables derivadas**

Se crea una variable "Antigüedad Meses" que es la antigüedad en meses calculada con la fórmula SIFECHA tomando como inicio "Member Since" y final "Contract end date C".

Dado que el Contract End Date es siempre final de mes, la antigüedad así calculada es correcta porque redondea hacia abajo.

Se crea otra variable por "Años" del cliente, que se calcula con la fórmula SIFECHA tomando como inicio "Birthday" y final la función HOY, en años.

Se crean dos variables más, año y mes de alta del Cliente, para poder hacer el segmetador por años y meses, respectivamente.

**4. Análisis de Patrones y detección de anomalías**

En el EDA se aprecian valores de edad muy elevados (+70 años)

En la hoja de Estadísicos\_Principales se analiza si se se trata de un outlier. Con una tabla dinámica se determina media, min, max, desviación típica y varianza.

Para determinar si el valor mínimo o máximo están contanimadas con la edad, se observa que la media y mediana son próximas y los valores extremos están dentro de las cajas con bigotes.

La mediana es 35 y la media es 36. Por tanto estos valores de la edad (70+) no se consideran outliers.

Por otro lado, he intentado realizar una Macro para que aparezcan las Altas, y otra macro para que aparezca la Facturación, pero no lo he conseguido: da un Error 1004, que es el mismo error que me sale al ejecutar el Excel de Netflix que se da como ejemplo.

Como conclusiones principales:

.-las altas de concentran entre los meses de septiembre-octubre-noviembre, y, en menor medida, en enero-febrero-marzo; esto, con independencia de la edad

.-los clientes se concentran prácticamente en el código postal del gimnasio, aunque hay un código postal conlidante, que debiera ser analizado (28918)

.-la edad se concentra en la franja de 18-32 sobre todo en hombres, y de 43-57 tanto hombres como mujeres

.-las cuotas se distribuyen bastante equilabradas entre individuales y familiares con 1 familiar