Análisis Churn Amazon

# PRE-ANALISIS:

* Idioma en Ingles -> Normalizar a Ingles todas las columnas
* Quito duplicados
* Marcar duplicados en rojo con formato condicional

OBSERVACIONES

* Hay que modificar la columna gender porque está en dos idiomas
* Comprobar under 30 como mayor de 30 y senior como mayor de 65 con una columna añadida y formula IF
* Datos de Payment method no están normalizados
* Cambiar tipo de dato monthly check mezcla números y números mas texto
* Customer segment -> Spanglish
* Average monthly dar formato decimal con dos decimales-> es moneda
* Numbers of complains muy importante habla de las quejas
* Applied discount -> Booleano aparece en castellano por la configuración regional
* Método de contacto puede ser importante para analizar cuál es la mejor manera de contacto con el cliente -> indicaría que hay problemas en algún departamento
* Contact date tiene errores de formato de fecha
* Datos incompletos (buscar si se pueden calcular)
* Churn category y churn reason importantes para abandono de clientes
* Contact date tiene errores de formato de fecha
* Customer ternure nos ha dicho Amazon que no es necesaria
* Average Monthly Expenses es en formato de dólares

# TRANSFORMACIÓN:

* Comparar valores de ChurnLabel y Churn puesto que la información es la misma
* Creamos columna nueva:
* =SI([@[Churn Label]]="No";[@Churned]=0;Y([@[Churn Label]]="Yes";[@Churned]=1;0;1))
* Vemos que la información es la misma por lo que eliminamos churn ya que, en este caso, nos interesa mas el formato texto que el formato booleano de 0 y 1 y la columna creada
* Eliminamos la columna teléfono porque para el análisis no es relevante
* Under 30 y senior permanecen porque es posible que sea una métrica de Amazon aunque aporta información derivada de age
* Eliminamos la columna Hobby, Favorite TV Show, internal notes(esta última no es relevante porque solo hay un dato)
* Transformar MONTHLY CHARGE a formato numero
* =VALOR.NUMERO(SI.ERROR(IZQUIERDA([@[Monthly Charge]];HALLAR(" ";[@[Monthly Charge]]));[@[Monthly Charge]]))
* opción 2:
* =SI.ERROR(extrae([@[Monthly Charge Sucio]];1;HALLAR(" ";[@[Monthly Charge Sucio]]))-1;[@[Monthly Charge Sucio]]) -> hace un -1 para extraer el espacio en blanco
* Contact date -> Formato de fecha
* =FECHANUMERO([@[Contact Date]]) -> devuelve número cambiamos a fecha corta
* Average Monthly Expenses Lo reconoce como formato general lo cambiamos a número con 2 decimales

NORMALIZACION

* Unlimited data plan
* Unlimited data plan cambiar 0 y 1 por no y yes
* =SI([@[Unlimited Data Plan]]=0;"No";"Yes")
* Gender cambiar Hombre a male
* =SI([@Gender]="Hombre";"Male";[@Gender])
* *opcion 2*

*Crear TABLA SUPLETORIA con los datos únicos de la columna original (y se puede automatizar con la función únicos y después crear columna nueva en tabla original y poner buscarv) y asignar valor para cada valor en castellano, con esto conseguimos que se actualice si se añaden valores nuevos*

* Payment method -> nompropio para poner todo en formato 1ª Mayusc y resto Minusc
* Customer segment , preferred method contact pasar a ingles
* applied discount -> Yes y No
* Volvemos a comprobar duplicados por si después de los cambios hay columnas nuevas duplicadas después de la transformación deseleccionando las columnas sucias (excepto senior que aún no sabemos la información correcta)
* elimina 317 nuevos duplicados.
* Quedan 3 valores duplicados por los valores de las columnas Monthly Charge(USD) y Total Charges.-> Verificar info con cliente.
* El cliente indica que los clientes mayores de 65 son senior, nos autoriza a cambiar el dato.
* Modificamos la columna senior según el valor de la columna AGE

# ANALISIS DESCRIPTIVO:

1. **\*Account Length (in months)**
   1. Media 32 (promedio)
   2. Mediana 29 (valor central)
   3. Moda 1 (el valor más repetido)
      1. Duración de cuentas es muy variada hay cuentas muy longevas y otras que solo duran un mes
   4. Desviación 24 (es una desviación muy alta)
   5. Curtosis -1.38 (negativa -> hay mas valores por debajo de la media
2. **\*Customer Service Calls**
   1. Desviación estandar (alta -> hay muchos que no llaman pero el que llama lo hace repetidamente
   2. Curtosis ( Hay mas peso en los extremos)
   3. Asimetría ( )
3. **\*Avg Monthly GB Download**
   1. Moda 0 (la mayor parte no descarga gb)
   2. Desviación (muy alta por lo que tenemos valores extremos)
   3. Curtosis ( es positiva -> las colas pesan mas -> la mayoría no descargan pero el que lo hace descarga mucho)
4. **\*Extra Data Charges**
   1. Desviación altísima (la moda es 0 y la median es cero -> valores más comunes, -> hay pocos con cargo extra pero los que tiene cargo es alto)

\*\* Buscar la relación entre las columnas anteriores

1. **\*Age**
   1. desviación muy alta (hay valores extremos que se alejan mucho de la media)
   2. curtosis (negativa porque los valores están más acumulados alrededor de la media)
   3. coeficiente ((positiva la mayor parte del grupo están por encima de la media -> la mayoría son mayores)
2. **\*Number of Customers in Group**
3. **\*Monthly Charge(USD)**
4. **\*Total Charges**
5. **\*Average Monthly Expenses**
6. **\*Number of Complaints or Support**