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

Los datos muestran que la duración de las cuentas es bastante variada con una media de 32 meses y una mediana de 29 meses pero con algunas cuentas que duran significativamente menos (un mes es la moda) y algunas que se extienden hasta 77 meses esto indica tanto una retención de clientes a largo plazo como una alta rotación de las cuentas más recientes

1. **\*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 ( )

los datos muestran que la mayoría de los clientes más de la mitad no realizan ninguna llamada al servicio de Atención al Cliente con una media de 0,92 llamadas sin embargo existe un pequeño número de clientes que realizan múltiples llamadas lo que sugiere que mientras muchos están satisfechos otros pueden enfrentar problemas que requieren atención adicional

1. **\*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)

Los datos muestran que aunque el promedio de descargas mensuales es de 6,69 GB hay una gran variabilidad en el uso de datos con muchos clientes descargando 0 GB y algunos alcanzando hasta 43 GB esto sugiere que el consumo de datos es diverso con un grupo significativo que utiliza el servicio de manera muy baja y otro que lo utiliza intensiva

1. **\*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

Los datos muestran que aunque el promedio de cargos adicionales es de 3,37 la mayoría de los clientes no incurren en esos costos ya que tanto la mediana como la moda son cero sin embargo hay un grupo significativo de clientes que enfrentan cargos altos lo que indica que existe una disparidad en el uso de datos que resulta en costos adicionales evidenciando tanto la eficiencia en la gestión de datos de muchos usuarios como la necesidad de atención a aquellos que incurrieron en altos cargos

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)

Los datos muestran que la edad promedio de los clientes es de aproximadamente 47 años con una mediana también de 47 años lo que sugiere que el grupo tiende a ser de mediana edad sin embargo hay un número significativo de clientes jóvenes como indica la moda de 29 años la amplia variabilidad de las edades reflejada en la alta desviación estándar y el rango destaca la diversidad del grupo de clientes lo que puede tener implicaciones en la segmentación del mercado y la estrategia de marketing de la empresa

1. **\*Number of Customers in Group**

los datos muestran que el promedio de clientes por grupo es de aproximadamente 0,84 lo que indica que la mayoría de los grupos no tiene clientes como lo sugieren la mediana y la moda aunque hay una gran variabilidad en el número de clientes por grupo con un máximo de 6 la alta desviación estándar y el coeficiente de asimetría indican que algunos grupos pueden tener una concentración inusualmente alta de clientes mientras que otros permanecen vacíos esto puede sugerir la necesidad de revisar cómo se están formando y gestionando estos grupos para optimizar la asignación de clientes

1. **\*Monthly Charge(USD)**

Los datos muestran que el cargo mensual promedio es de aproximadamente 30,93 USD con una mediana también de 31 USB lo que indica una distribución equilibrada, sin embargo la moda de 10 USD, sugiere que muchos clientes optan por planes más económicos. La amplia variabilidad en los cargos evidenciada por la alta desviación estándar y el rango, resalta que existen diferentes niveles de servicio lo que puede ser estratégico para atender a diversos segmentos de mercado

1. **\*Total Charges**

Los datos muestran que el cargo total promedio es de aproximadamente 1085,43 USD con una mediana de 651 USD que indica que la mitad de los clientes tiene un gasto total relativamente bajo, la alta desviación estándar y varianza reflejan una gran diversidad en los totales acumulados, lo que sugiere que algunos clientes han estado con el servicio por mucho tiempo y han acumulado gastos altos mientras que otros han tenido una relación más breve y menos costosa con la empresa. Esto podría indicar la necesidad de segmentar estrategias de marketing y Atención al Cliente para abordar las diferentes necesidades de estos grupos

1. **\*Average Monthly Expenses**

En resumen coma los datos muestran que el gasto mensual promedio de los clientes es de aproximadamente 31 dólares, con una mediana cercana que sugiere que muchos clientes se agrupan en torno a este valor. Sin embargo, la moda más alta de 47,18 indica que existe un grupo significativo de clientes que gasta más. La alta desviación estándar y varianza reflejan una gran diversidad en los hábitos de gasto, lo que puede ser útil para la segmentación y el diseño de ofertas personalizadas

1. **\*Number of Complaints or Support**