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:

## NUMERICAS

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**

## CATEGORICAS Y TEMPORALES

Las columnas categóricas que analizamos son :

* **Unlimited Data Plan**
  + Los clientes tienden a abandonar el servicio si han contratado los datos ilimitados, ya que pasan de una tasa de abandono del 16,12% a 32,11%. esto puede deberse a que no sea un servicio acorde a las necesidades de los usuarios y esto genere insatisfacción en los mismos.
* **State**
* **Gender**
  + Parece que no es un factor determinante en la redención o el abandono del servicio ya que todos los datos giran en torno al 27% de abandono. Sin embargo es cierto que aumenta ligeramente en el caso de los usuarios que prefieren no compartir su información de género (28,57%)
* **Under 30**
  + Los usuarios de menos de 30 tienen una mayor fidelidad con el servicio ya que estos tienen una tasa de abandono del 22,98% frente al 27,79% de los mayores de 3. Esto puede deberse a un contenido más enfocado a un público joven, o una mayor facilidad a la hora de interactuar con la plataforma
* **Senior**
  + Los usuarios de la categoría sénior (más 65) tienden a abandonar el servicio (38,46%) frente a los menores de 65 que tienen una tasa de abandono del 24,18%. Esto puede estar relacionado con la columna anterior habiendo problemas con la accesibilidad de la plataforma o contenido no acorde con los gustos de este sector de clientes
* **Group**
  + Los usuarios pertenecientes a un pack grupo tienen una tasa de abandono de solo 6,5%. Sugiere una gran satisfacción por parte de los clientes que pertenecen a este sector. Probablemente porque el coste del usuario sea menor. Recomendaríamos al cliente potenciar los planes de grupo
* **Device Protection & Online Backup**
  + Los usuarios que pagan por este servicio es más fácil que sigan con nosotros ya que la tasa de abandono pasa del 28,87% al 23,02%. Puede ser que les ofrezcamos seguridad a la hora de no perder sus fotos y vídeos, en resumen, sus recuerdos. Además es posible que los clientes que contrataron este servicio lo hayan hecho después de un ejercicio de proyección a futuro haciendo un análisis previo y decidiendo que con nosotros era con los que iban a mantener sus fotos y por ello no se quieran ir del servicio
* **Contract Type**
  + la permanencia a la plataforma parece que está estrechamente ligada a la duración del contrato ya que podemos observar cómo los contratos mes a mes tienen una tasa de abandono del 46,31%, año a año del 11,28% y de 2 años solo del 2,78%. Esto puede deberse a que muchos usuarios solo contratan la opción de un mes por varias razones periodo de prueba, compras estacionales, estreno de series concretas, etc. Sería conveniente potenciar la venta de los contratos de larga duración.
* **Payment Method**
  + Tenemos nulos. (16 usuarios). Hay una diferencia abrupta entre el abandono de los usuarios que pagan con paper check 38,21%, sin embargo el número de usuarios que utiliza este método es muy pequeño respecto a los otros dos métodos de pago. Adicionalmente, observamos como los usuarios que pagaron con débito son más propensos a abandonar el servicio 34,53%. Esto puede deberse a algún problema con la plataforma de pago. Además habría que estudiar el poder adquisitivo de los usuarios que tienen tarjeta de crédito frente a tarjeta de de
* **Churn Category**
  + No hemos hecho un análisis comparativo de la categoría de abandono con respecto a si el cliente ha abandonado o no ya que no tenía sentido esa tabla. Sin embargo hemos obtenido que hay clientes que abandonan pero que no nos indican la razón con un porcentaje muy pequeño 0,55% y además hemos estudiado que la razón principal del abandono de nuestros clientes es por la competencia
* **Churn Reason**
* **Customer Segment**
  + Podemos observar cómo el segmento del cliente no es determinante para identificar algún patrón en el abandono o no de la plataforma, ya que todas las categorías tienen una tasa de abandono cercana al 26%
* **Preferred Contact Method**
  + Al igual que en la columna anterior no parece que haya una relación directa entre el abandono y el método de contar
* **Applied Discount**
  + aunque pudiera parecer en un momento lógico que los clientes decidieran quedarse en el servicio si les aplicábamos un descuento los datos nos han demostrado que no es una característica que este relacionada con el abandono

Las columnas temporales que analizamos son :

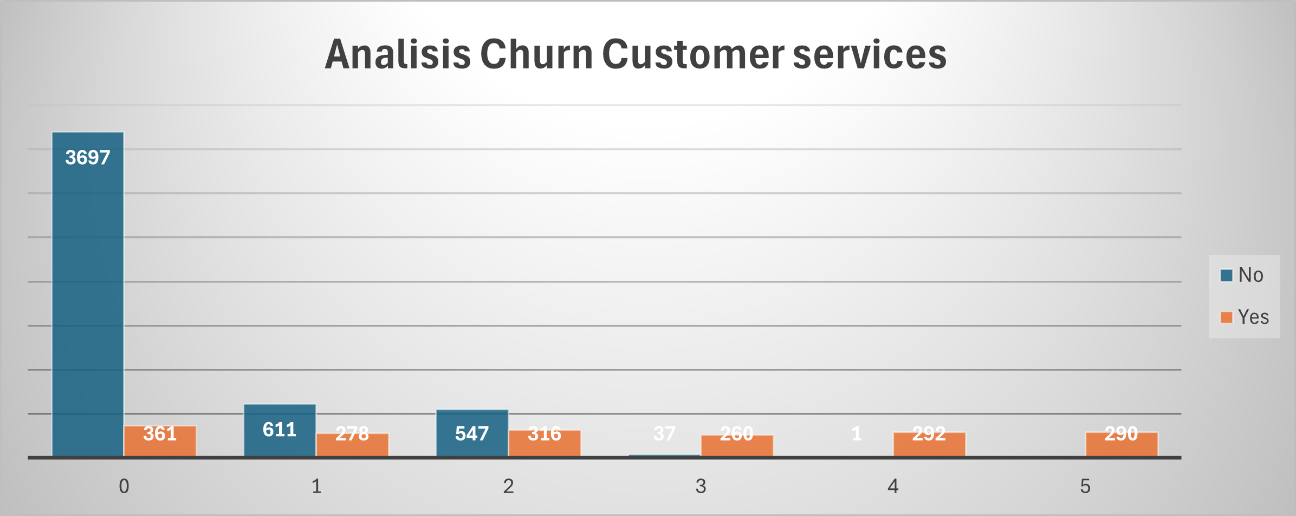
* **Contact Date**
  + **Año:** pudimos observar que el año de contacto donde más tasa de abandono hubo fue 2022 con un 28,43 por 100 mientras que el que menos fue 2020,1 25,47%
  + **Trimestre:** en este caso vimos cómo el trimestre de contacto con mayor tasa de abandono es el segundo abril junio esto puede ser debido a que los usuarios que contrataron este servicio lo hicieron para utilizarlo durante el periodo vacacional pero no les interesa tenerlo durante todo el año
  + **Mes:** en este caso pudimos ver cómo el mes donde se realizó el contacto con mayor abandono es mayo con un 29,55% acorde a los datos trimestrales
* **Last Transaction Date**
  + **Año:** Podemos ver cómo el año en el que ha habido una mayor tasa de abandono es el 2022,1 27,28%. Esto pudo deberse a que a partir de ese año las restricciones por COVID en EEUU prácticamente desaparecieron, por lo tanto, menos gente decidiera quedarse en casa y no fuera necesario seguir con nuestro servicio
  + **Trimestre:** En este caso el trimestre con mayor tasa de abandono es el primero o en un 28,77% esto puede deberse a que los usuarios no quieren seguir con la suscripción para el segundo trimestre ya que al empezar la primavera pasan más tiempo fuera de casa
  + **Año:** sin embargo podemos observar cómo el mes con mayor tasa de abandono es agosto muy probablemente debido a que haya mucha gente que se vaya de vacaciones ese mes y no vaya a disfrutar del servicio durante el periodo vacacional prefiriendo darse de baja en agosto para volver a suscribirse al mes siguiente.

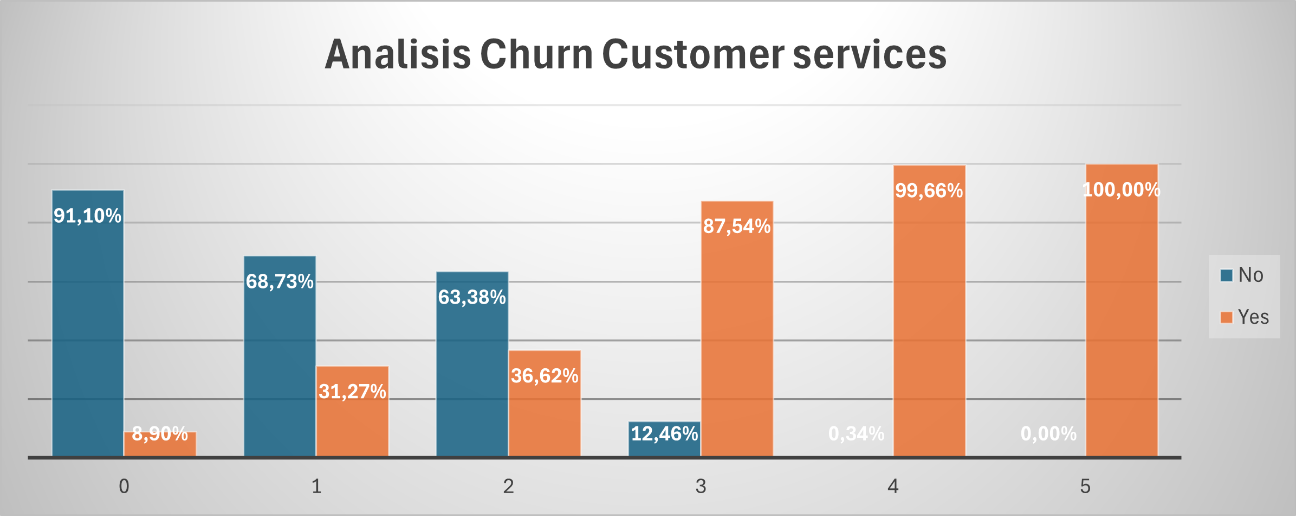
## Insights

### Numéricas

* **Customer Service Calls**

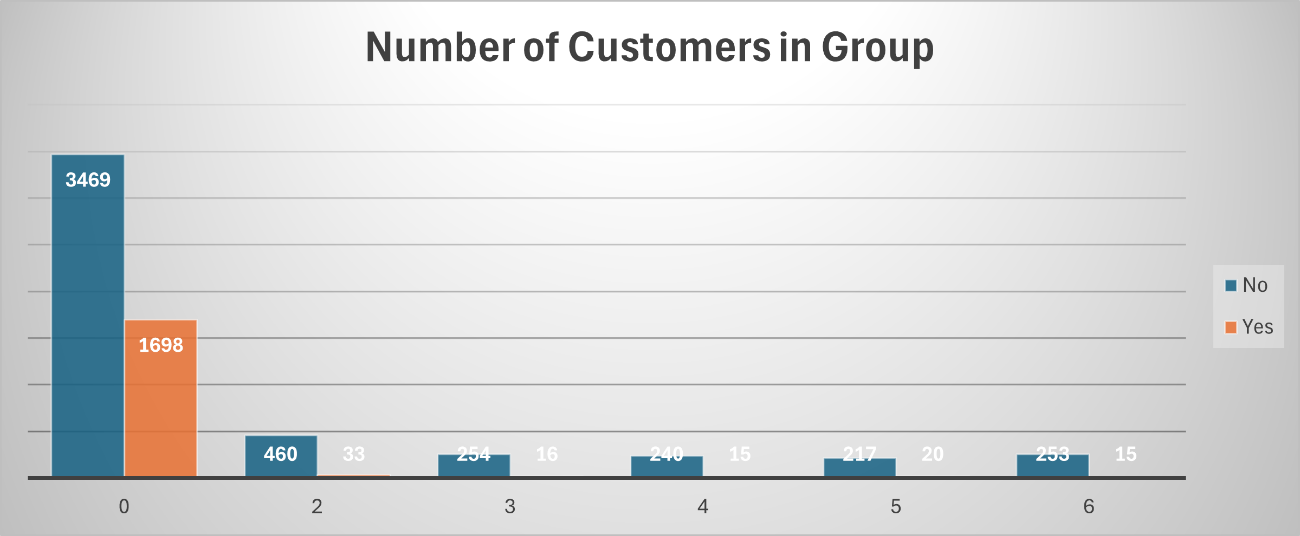
Vemos un incremento tanto a nivel absoluto como porcentual por fila de nuestra tabla dinámica del número de personas que abandonan el servicio según aumenta el número de llamadas al servicio de customer. Esta columna es relevante para nuestro análisis

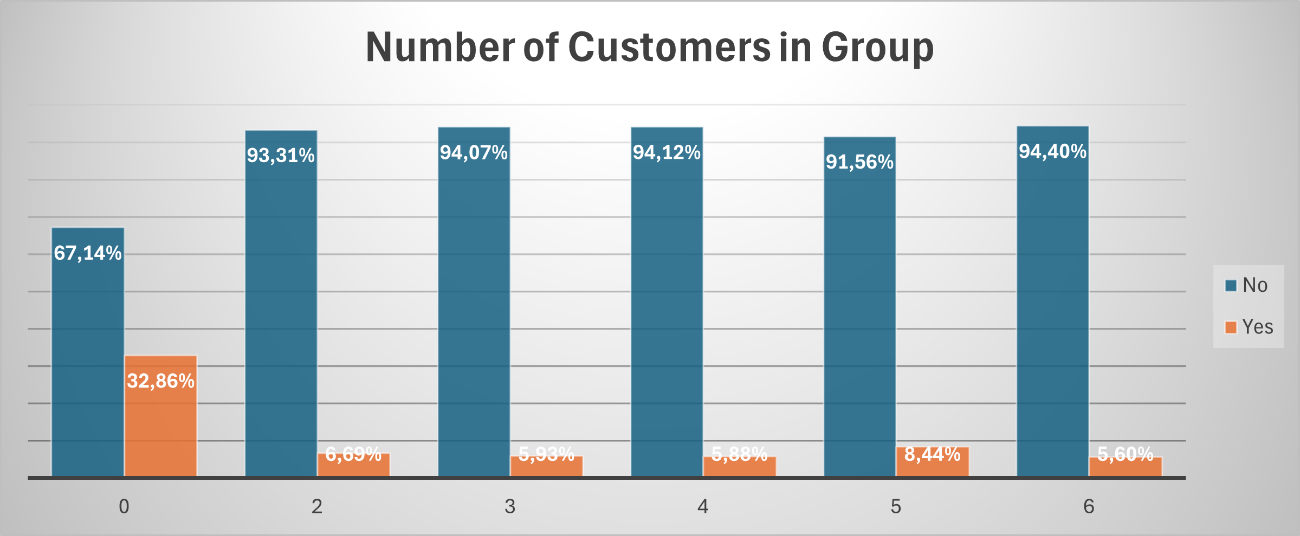




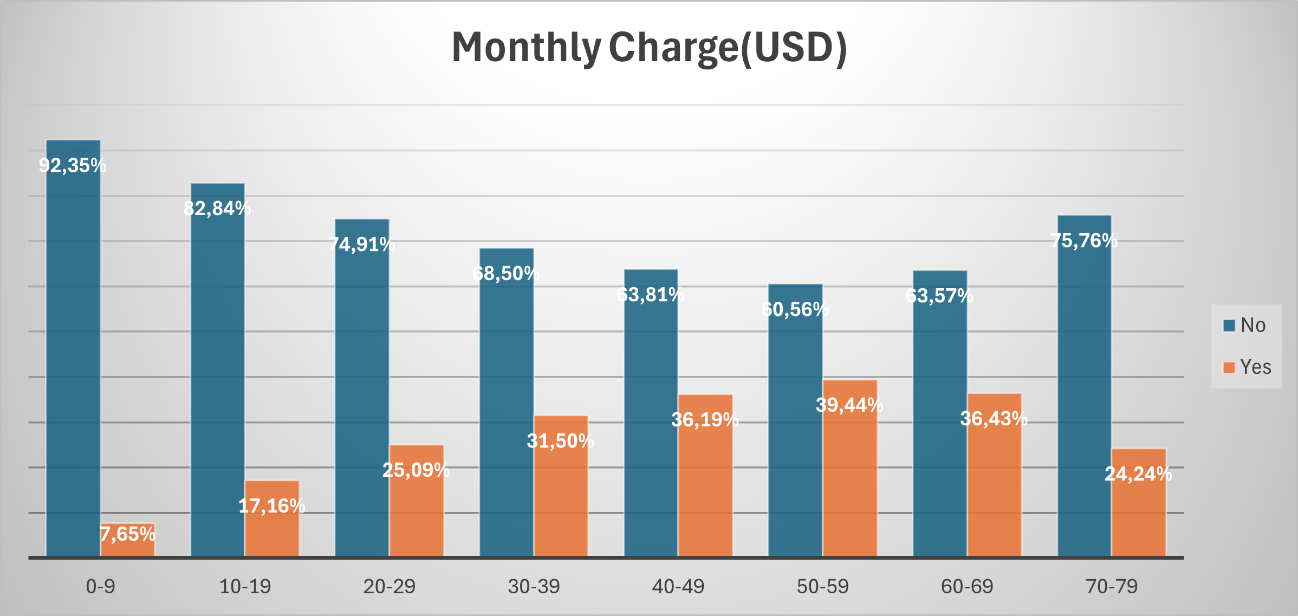
* **Number of Customers in Group**

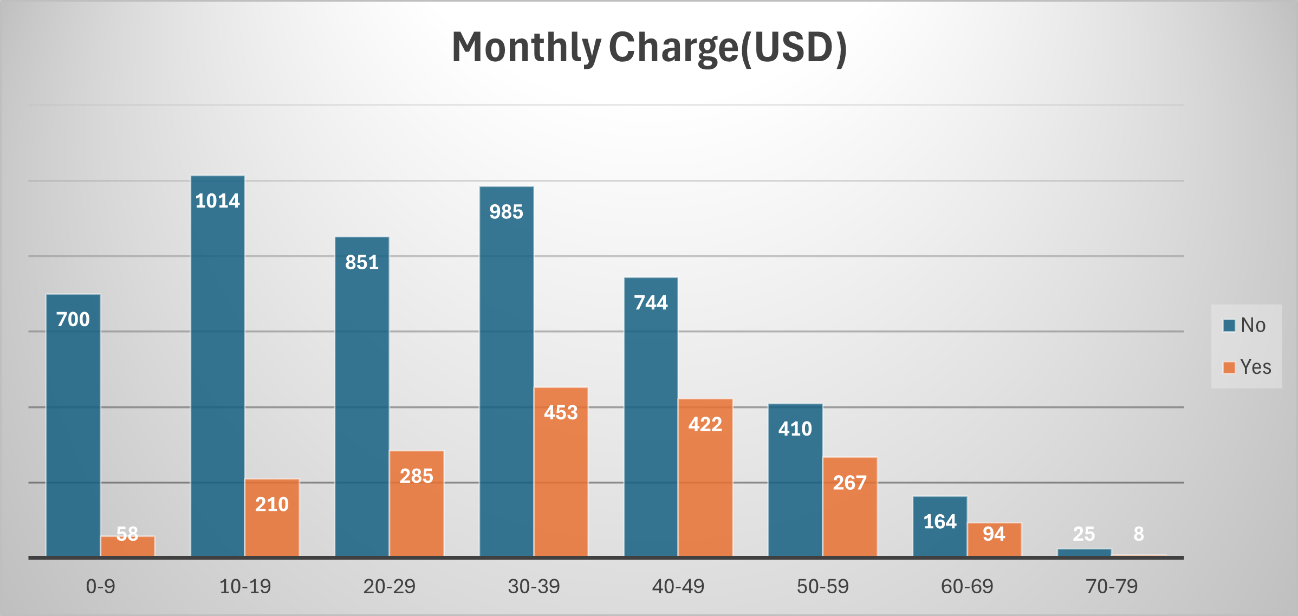
Vemos que la mayor parte de las personas no tienen contratado un plan familiar. Si lo relacionamos con el Churn vemos que la mayor parte de los clientes que abandonan el servicio no tienen contratado un plan familiar mientras que la tasa de abandono en los clientes con varias personas en un plan familiar tiene una tasa de abandono muy baja





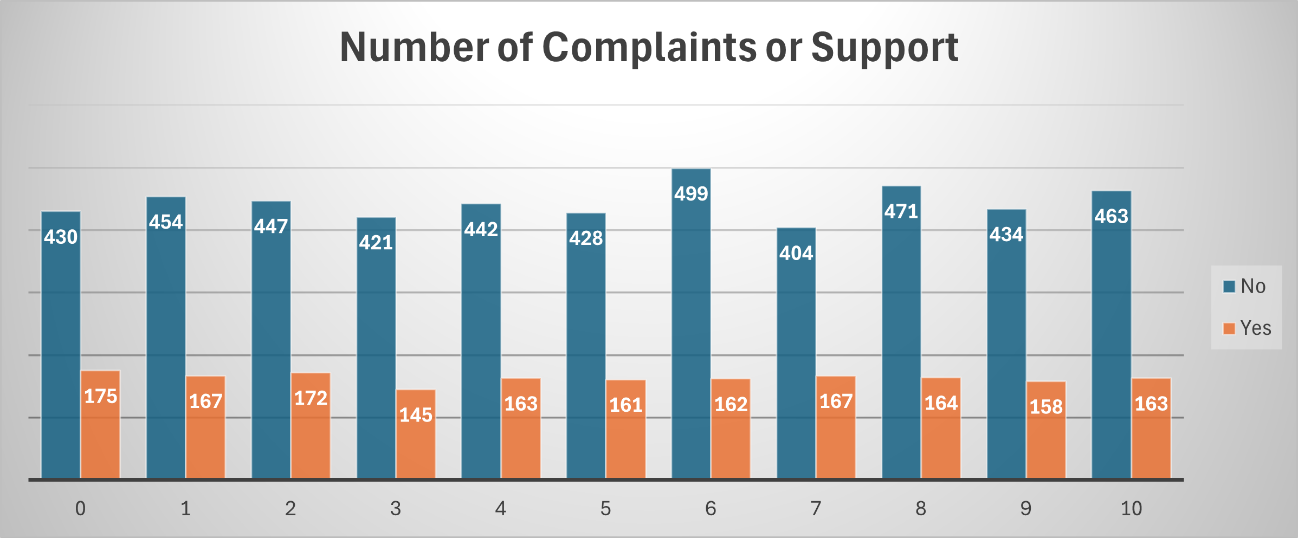
* **Monthly Charge(USD)**

Vemos que hay un aumento en el número de abandonos en términos porcentuales según aumentan los cargos mensuales disminuyendo a partir de los 60$

Sin embargo si vemos los términos absolutos esta bajada a partir de los 60$ puede deberse a la disminución de la cantidad de datos

**ANÁNALISIS ADICIONALES**

Como análisis adicionales podemos ver la columna number of complaints or support. A priori pensábamos que podía ser una columna con valores reseñables para entender la tasa de abandono pero hemos visto que los datos se distribuyen de manera uniforme independientemente del número de contactos a support.



En cuanto a la columna age podemos ver que la mayoría parte de los datos los tenemos entre los 19 y los 68 años. Sin embargo se observa un aumento en la tasa de abandono a partir de los 69 años, lo que podría ser relevante para nuestro análisis

