# 

# Tabla de contenido

[**Tabla de contenido**](#_30j0zll) **2**

[**Descripción del caso de negocio**](#_c7919pissta2) **3**

[**Objetivos del trabajo**](#_2et92p0) **3**

[**Datos**](#_tyjcwt) **4**

[**Manipulación de datos**](#_3dy6vkm) **6**

[**Hallazgos encontrados por el EDA**](#_1t3h5sf) **8**

[Análisis Univariado](#_4d34og8) 8

[Análisis Bivariado](#_2s8eyo1) 12

[Variable Churn Vs. Numéricas](#_17dp8vu) 12

[Variable Churn Vs. Categóricas](#_26in1rg) 15

[Numéricas Vs. Numéricas](#_35nkun2) 19

[Categóricas Vs. Numéricas](#_44sinio) 21

[Categóricas Vs. Categóricas](#_2jxsxqh) 28

[Análisis Multivariado](#_z337ya) 28

[Conclusion](#_cf1tnnxkvzi) 36

[**Desarrollo del Modelo**](#_1xf9opvw08c7) **37**

[Random Forest](#_z5lo1e7g1c8p) 37

[Regresion Logistica:](#_68uijsaebsr4) 37

[Metricas](#_ijj59ee1zes2) 37

# 

# 

# Descripción del caso de negocio

Las empresas cuyo modelo de negocio refiere a la prestación de un servicio, tienen como principal objetivo lograr captar y retener clientes constantemente. Sin dudas que este es un desafío difícil teniendo en cuenta el acceso que tienen los consumidores a información de todo tipo; como sus derechos de usuario, funcionamiento del servicio, quejas o reclamos, innovación de la empresa o empresas competidoras, precios, etc.

Resulta mucho más costoso captar nuevos clientes que retenerlos, debido a lo costosas que son las campañas de marketing que ejecutan las empresas para incrementar su volúmen de ventas. Por eso es importante contar con indicadores que permitan medir, controlar, y tomar acciones preventivas frente a posibles abandonos de servicio de parte de los consumidores.

La predicción de estos sucesos es, probablemente, uno de los casos de aplicación más importantes de la analítica de datos en el sector comercial. *Churn* hace referencia, en términos comerciales, al hecho de que un cliente cancele una suscripción a un servicio que ha estado usando. Un ejemplo común es la gente que cancela su suscripción a Spotify o Netflix. Por lo tanto, dicha predicción consiste esencialmente en qué clientes tienen más probabilidades de cancelar una suscripción.

En el presente trabajo abordaremos esta temática, tomando como base un dataset extraído de IBM, que cuenta con información de una cartera de clientes perteneciente a una empresa de telecomunicaciones.

# Objetivos del trabajo

El objetivo del trabajo es predecir qué clientes tienen altas probabilidades de abandonar algún tipo de sus servicios mediante la aplicación de un algoritmo de clasificación.

Para poder llegar a este objetivo realizaremos, en primer lugar, un análisis preliminar y exhaustivo de las distintas variables presentes en nuestro dataset para descubrir y entender su comportamiento (EDA). Esto nos permitirá tomar decisiones relacionadas a la limpieza de datos y selección de variables, para luego poder aplicar los modelos.

Mediante el EDA se intentará encontrar tendencias, patrones y relaciones entre variables; desarrollándose tres análisis distintos denominados univariado, bivariado y multivariado. El univariado explora el comportamiento de cada variable de forma individual. El bivariado se corresponde con el análisis de la relación entre dos variables, mientras que el multivariado es necesario cuando se deben analizar más de dos variables en forma simultánea. Cabe aclarar que los análisis anteriores corresponden tanto a variables numéricas como categóricas.

Una vez concluído el análisis exploratorio de datos, utilizaremos distintos algoritmos de clasificación para predecir la variable *Churn*. Realizaremos la selección del algoritmo en base a distintas métricas de performance que nos indicarán qué modelo predice mejor el abandono de clientes.

# Datos

A continuación se presenta un listado con las variables que analizaremos a lo largo del proyecto, trabajaremos utilizando el lenguaje de programación Python.

La forma en que se almacena la información en un DataFrame u objeto Python afecta a lo que podemos hacer con él y también a los resultados de los cálculos.

Hay dos tipos principales de datos que estaremos explorando en este trabajo: numéricos y de texto. Los tipos de datos numéricos incluyen enteros (integer) y números de punto flotante (float). Un número de punto flotante tiene puntos decimales incluso si el valor del punto decimal es 0. Un integer nunca tendrá un punto decimal. Así que, si quisiéramos almacenar 1.13 como un entero de tipo integer se almacenará como 1. En Python se ve el tipo de dato Int64 que representa un entero de 64 bits. Por último, el tipo de datos de texto se conoce como secuencia de caracteres (string). En Pandas se los conoce como objetos (object). Las secuencias de caracteres pueden contener números y / o caracteres.

***Object (28)***

1. Customer\_id: ID del Cliente.
2. Interaction: Identificaciones únicas relacionadas con transacciones de clientes, soporte técnico e inscripciones.
3. City: Ciudad de residencia del cliente que figura en el estado de cuenta.
4. State: Estado de residencia del cliente como se indica en el estado de cuenta.
5. County: Condado de residencia del cliente como se indica en el estado de cuenta.
6. Area: Tipo de área (rural, urbana, suburbana), según datos del censo
7. Timezone: Zona horaria de residencia del cliente basada en la información de registro del cliente.
8. Job: Ocupación del cliente como se indica en la información de registro
9. Education: Grado de educación más alto obtenido por el cliente según lo declarado en la información de registro.
10. Employment: Estado de empleo del cliente según lo declarado en la información de registro.
11. Marital: Estado civil del cliente según lo indicado en la información de registro.
12. Gender: Autopercepción del cliente como hombre, mujer o no binario.
13. Churn: : El cliente interrumpió el servicio en el último mes (sí/no).
14. Techie: El cliente se considera tecnológico (según el cuestionario del cliente cuando se inscribió en los servicios) (sí/no).
15. Contract: Plazo del contrato del cliente (mes a mes, un año, dos años).
16. Port\_modem: El cliente tiene un módem portátil (sí/no).
17. Tablet: El cliente posee una tableta como iPad, Surface, etc. (sí/no).
18. InternetService: Proveedor de servicios de Internet del cliente (DSL, fibra óptica, Ninguno).
19. Phone: El cliente tiene servicio telefónico (sí/no).
20. Multiple: El cliente tiene varias líneas (sí/no).
21. OnlineSecurity: El cliente tiene un complemento de seguridad online (sí/no).
22. OnlineBackup: El cliente tiene un complemento de copia de seguridad online (sí/no).
23. DeviceProtection: El cliente tiene un complemento de protección de dispositivos (sí/no).
24. TechSupport: El cliente tiene un complemento de soporte técnico (sí/no).
25. StreamingTV: El cliente tiene servicio de transmisión de TV (sí/no).
26. StreamingMovies: El cliente tiene películas en streaming (sí/no).
27. PaperlessBilling: El cliente tiene facturación electrónica (sí/no).
28. PaymentMethod: Método de pago del cliente (cheque electrónico, cheque enviado por correo, banco (transferencia bancaria automática), tarjeta de crédito (automática)).

***Int64 (14)***

1. CaseOrder: Variable de marcador de posición para conservar el orden original del archivo de datos sin procesar.
2. Zip: Código Postal de residencia del cliente que figura en el estado de cuenta.
3. Population: Población dentro de un radio de una milla del cliente, según datos del censo.
4. Email: Cantidad de correos electrónicos enviados al cliente en el último año (marketing o correspondencia).
5. Contacts: Número de veces que el cliente se comunicó con el soporte técnico.
6. Yearly\_equip\_failure: Cantidad de veces que el equipo del cliente falló y tuvo que reiniciarse/reemplazarse en el último año.

Las siguientes variables, representan las respuestas a una encuesta de ocho preguntas, en las que se pide a los clientes que califiquen la importancia de varios factores en una escala del 1 al 8 (1 = más importante, 8 = menos importante):

1. Item1: Respuesta oportuna
2. Item2: Reparaciones oportunas
3. Item3: Reemplazos oportunos
4. Item4: Fiabilidad
5. Item5: Opciones
6. Item6: Respuesta respetuosa
7. Item7: Intercambio cortés
8. Item8: Evidencia de escucha activa

***Float64 (9)***

1. Lat: Coordenada de latitud de la residencia del cliente (GPS), que figura en el estado de cuenta.
2. Lng: Coordenada de longitud de la residencia del cliente (GPS), que figura en el estado de cuenta.
3. Children: Cantidad de niños en el hogar del cliente según lo informado en el registro.
4. Age: Edad del cliente según lo informado en el registro
5. Income: Ingreso anual del cliente según lo informado en el momento del registro.
6. Outage\_sec\_perweek: Promedio de segundos por semana de interrupciones del sistema en el vecindario del cliente.
7. Tenure: Número de meses que el cliente se ha quedado con el proveedor.
8. MonthlyCharge: Importe cobrado al cliente mensualmente. Este valor refleja un promedio por cliente.
9. Bandwidth\_GB\_Year: Cantidad promedio de datos utilizados (GB), en un año por el cliente.

# Manipulación de datos

En esta sección analizaremos la existencia de valores nulos y repetidos en todas las columnas, así como el tipo de dato que representa cada una, y comenzamos a visibilizar probables transformaciones en nuestro conjunto de datos.

* churn\_bool: Creamos una variable booleana de churn con 0 y 1 para poder compararla contra variables numéricas
* total\_encuesta: Variable que unifica los valores obtenidos en la encuesta( item 1 al 8)
* Cambiamos los nombres de las últimas columnas reemplazando itemx por el nombre del item:

'item1': 'timely\_response',

'item2': 'timely\_fixes',

'item3': 'timely\_replacements',

‘item4': 'reliability',

'item5': 'options',

'item6': 'respectful\_response',

'item7': 'courteous\_exchange',

'item8': 'active\_listening'

* Creamos nuevas variables que cuentan la cantidad de cada tipo de servicios y el total de servicios que los clientes contratan:
  + q\_online\_serv: cuenta si el cliente tiene contratado internet service, online backup, onlinesecurity y/o tech support. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 4.
  + q\_phone\_serv: cuenta si el cliente tiene contratado phone y/o device protection. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 2.
  + q\_streaming: cuenta si el cliente tiene contratado streaming tv y/o streaming movies. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 2.
  + q\_total\_serv: realiza una suma de las tres categorías nombradas anteriormente siendo el valor máximo de esta variable = 8.
* zip\_zone: creación de variable geográfica agrupando los zip por su primer dígito. Este número indica una de las 9 zonas geográficas generales de EEUU.
* Creamos variables string que se correspondan con las numéricas categóricas para poder realizar análisis de variables categóricas:
  + children\_cat, age\_cat, zip\_cat pop\_cat, email\_cat, contacts\_cat, failure\_cat, timely\_response\_cat, timely\_fixes\_cat, timely\_replacements\_cat, reliability\_cat options\_cat, respectful\_response\_cat, courteous\_exchange\_cat, active\_listening\_cat, total\_encuesta\_cat, q\_online\_serv\_cat, q\_phone\_serv\_cat, q\_total\_serv\_cat, q\_streaming\_cat, zip\_zone\_cat.

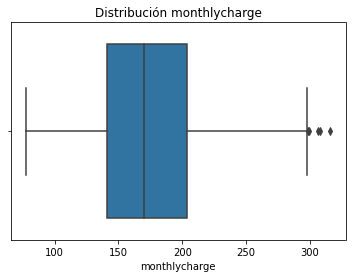
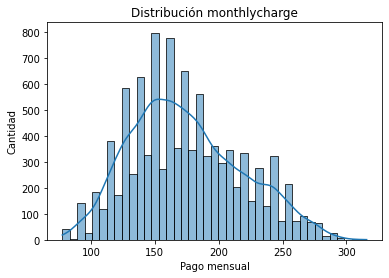
# Hallazgos encontrados por el EDA

## Análisis Univariado

Como mencionamos anteriormente este análisis consiste en explorar cada variable por separado, individualmente. Además, intentaremos determinar la distribución de algunas de las variables del proyecto, y en caso no presentar una distribución normal, aplicarles una transformación para lograr su correcta manipulación. Como regla general, se trata de escoger una transformación que conduzca a una distribución simétrica, y más cercana a la distribución normal, ya que si la variable es normal quiere decir que se pueden realizar estimaciones de ella en base a una función que depende de la media y el desvío. De este modo, se pueden aplicar numerosas técnicas de inferencia estadística. En nuestro caso, evaluaremos la transformación de variables aplicándoles logaritmo.

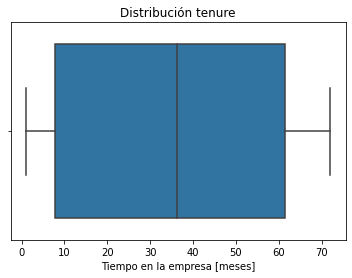
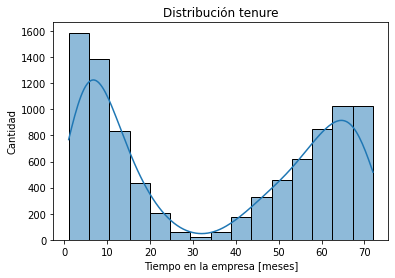
En esta sección serán importantes los conceptos de distribución normal, curtosis y valor atípico. La curtosis de una variable estadística/aleatoria es una característica de forma de su distribución de frecuencias/probabilidad. Una curtosis grande implica una mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios. Esto explica una forma de la distribución de frecuencias/probabilidad con colas más gruesas, con un centro más apuntado y una menor proporción de valores intermedios entre el pico y las colas. En segundo lugar, la distribución normal es un modelo teórico capaz de aproximar satisfactoriamente el valor de una [variable aleatoria](https://economipedia.com/definiciones/variable-aleatoria.html) a una situación ideal. En otras palabras, la distribución normal adapta una variable aleatoria a una función que depende de la [media](https://economipedia.com/definiciones/media.html) y la [desviación típica.](https://economipedia.com/definiciones/desviacion-tipica.html) Es decir, la [función](https://economipedia.com/definiciones/funcion-matematica.html) y la variable aleatoria tendrán la misma representación pero con ligeras diferencias. Por último, un valor atípico es una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. Los valores atípicos suelen ser datos que pertenecen a una población diferente del resto de las muestras establecidas.

***Monthly Charge***

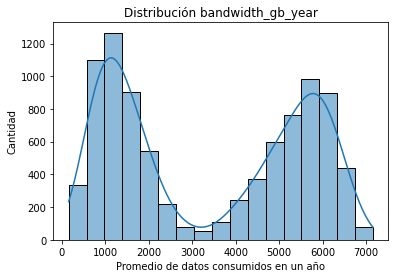
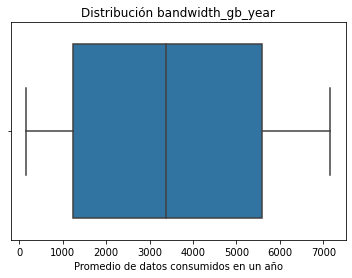


Se observa una variable relativamente normalizada, es decir, se aproxima a una normal, por lo que no hará falta aplicarle una transformación. Al ser la distribución normal una distribución simétrica, el valor de la moda, media y mediana, deben coincidir. En otras palabras, los valores más frecuentes deben estar alrededor de la media, y esto no sucede en su totalidad. Se observa en el histograma que la mayor cantidad de datos se encuentra concentrada en el orden de los 150 dólares mensuales, es decir, estamos en presencia de una distribución asimétrica hacia la izquierda y esto implica que sus sesgo es negativo. Además, los clientes que más aportan a los beneficios de la empresa pagan aproximadamente 300 dólares mensuales, mientras que los que menos aportan pagan aproximadamente 80 dólares. Por último, cabe recalcar que existen valores atípicos en el orden de los 300 dólares mensuales.

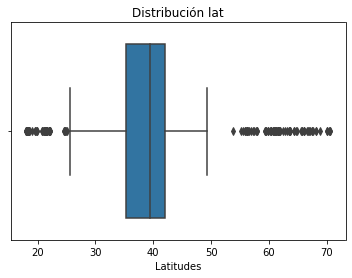
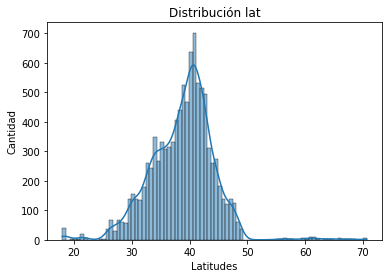
***Tenure***

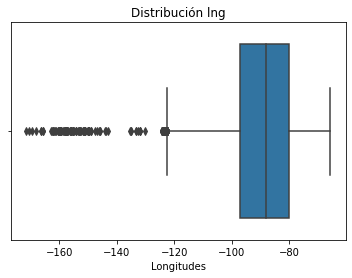


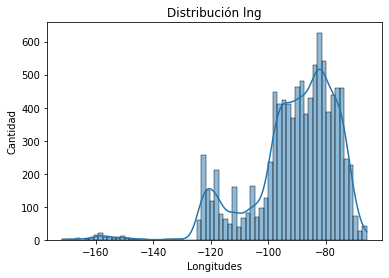
El histograma permite ver dos picos que representan distintas modas, es decir, los dos valores más comunes en nuestro conjunto de datos. Esto quiere decir que la variable sigue una distribución bimodal, lo que implica que se distinguen dos tipos de clientes principales: quienes están hace más de 40 meses y quienes no superan los 20 meses de antigüedad. Ahora bien, a la hora de analizar el porqué de esta distribución, podemos pensar en que la gente decide rápidamente si quiere conservar el servicio o rotar hacia la competencia. Es sabido que las empresas de telecomunicaciones se “tironean” clientes constantemente, por lo que hace sentido que la distribución sea así. Una mejor manera de analizar e interpretar las distribuciones bimodales es simplemente dividir los datos en dos grupos separados, para luego analizar el centro y la dispersión de cada grupo, pero no es algo que nos interese en este proyecto.

***Bandwidth\_Gb\_Year***

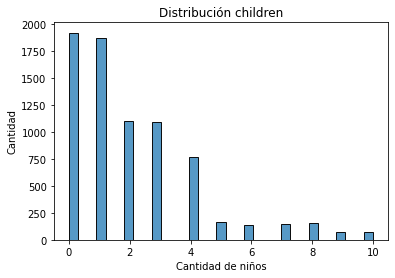
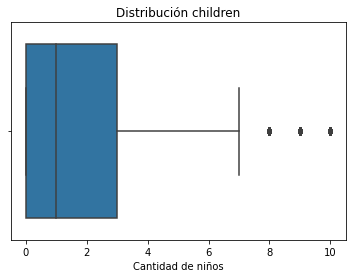
Esta variable también presenta una distribución bimodal, lo que implica que se distinguen dos grupos de clientes principales: quienes consumen anualmente alrededor de 5500 GB, y quienes consumen 1000 GB aproximadamente.

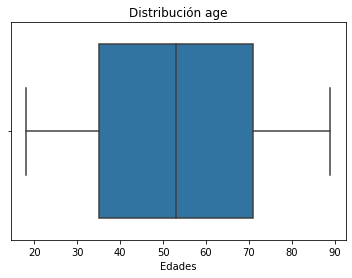
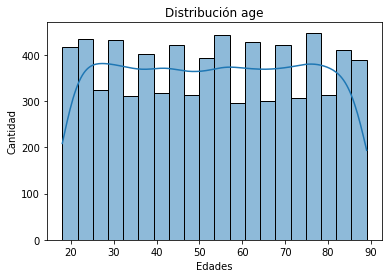
***Lat***

***Lng***

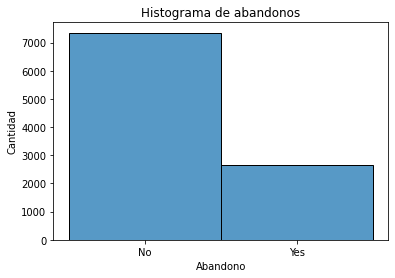
******

*Lat* y *Lng* son variables que nos permitirán saber la ubicación de los clientes. Latitud tiene una distribución más próxima a la normal y con menos dispersión que longitud, que tiene un sesgo hacia la derecha. Esto podría indicar que es un país o zona con mayor territorio horizontal que vertical, por lo que podemos intuir que la gran mayoría de los clientes son de US. También presentan muchos outliers.

***Children***

***Age***

Por un lado, children es una variable entera que presenta muchos outliers. La mayoría de ellos representan gente con muchos hijos, por lo que se podría considerar armar una categoría (más de tantos hijos). Además se ve que la gran mayoría de los clientes no tienen hijos. Age, por otro lado, tiene una distribución más bien uniforme, es decir, tenemos una cartera de clientes de todas las edades. Ambas variables son numéricas categóricas.

***Churn***

Hay un 26,5% de clientes que abandonan la empresa, mientras que un 73,5% continúan con el servicio. Al ser *churn* la variable target, consideramos que el dataset está al límite de estar desbalanceado. Un conjunto de datos desbalanceado es uno donde el número de observaciones pertenecientes a un grupo o clase es significativamente mayor que las pertenecientes a las otras clases. Por esta razón es que se calcularán las métricas con el dataset balanceado y se determinará con cuál de ellos trabajar.

En relación a las otras variables categóricas, hay unas cuestiones a tener en cuenta. Las variables customer\_id, interaction, city, county, state y timezone tienen muchos valores distintos. De las variables con metadato, consideraríamos trabajar únicamente con state. Job y education tienen muchas categorías pero se podrían disminuir al realizar agrupaciones. En adición, vimos que la mayoría de los contratos son mes a mes.

## Análisis Bivariado

En términos generales, el análisis bivariado es la investigación de la relación entre dos conjuntos de datos,como pares de observaciones tomadas de una misma muestra o individuo.

Análisis de relaciones con la variable objetivo: **Churn**

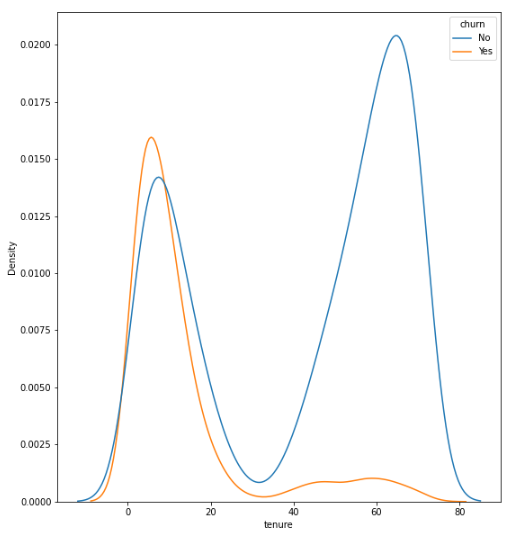
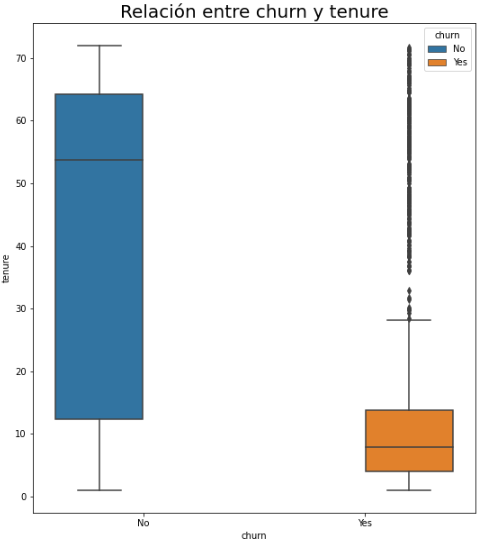
Decidimos hacer primero el análisis de las relaciones entre la variable objetivo y el resto de las variables, ya que esta es la variable principal de nuestro proyecto y consideramos que debe tener un tratamiento especial en el análisis.

### Variable Churn Vs. Numéricas

Para empezar con este análisis, creamos la variable churn\_bool para poder tomar a churn cómo numérica, reemplazando los “Yes” por 1 y los “No” por 0.

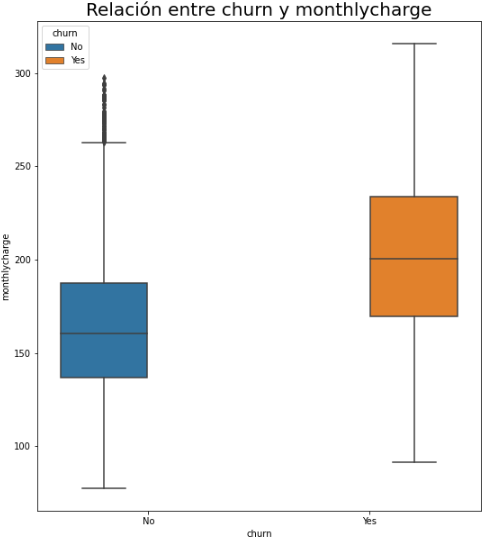
Luego, analizamos las relaciones entre churn\_bool y todas las variables numéricas, quedándonos con aquellas que presentan comportamientos marcadamente diferentes contra no churn.

**Churn Vs. Tenure**



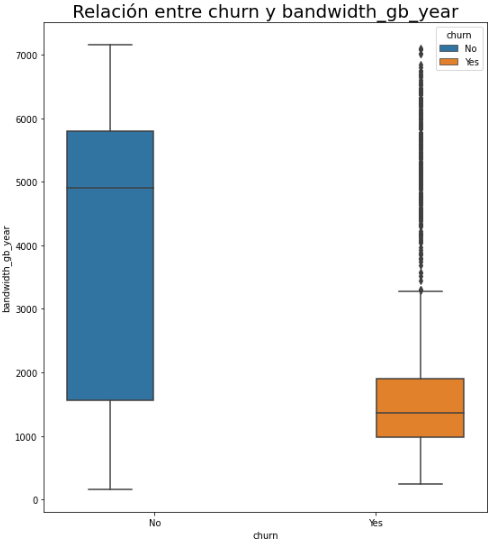
La mayoría de los churn se concentran en personas con tenure más baja de 30 meses. Hay outliers en los clientes más antiguos, pero en general estos se quedan. La mediana de los churn y no churn es muy distinta, al igual que su dispersión. Churn tiene un sesgo a la derecha concentrando valores en antigüedad baja pero con una cola importante en valores de tenure alta.

**Churn Vs. Monthly Charge**



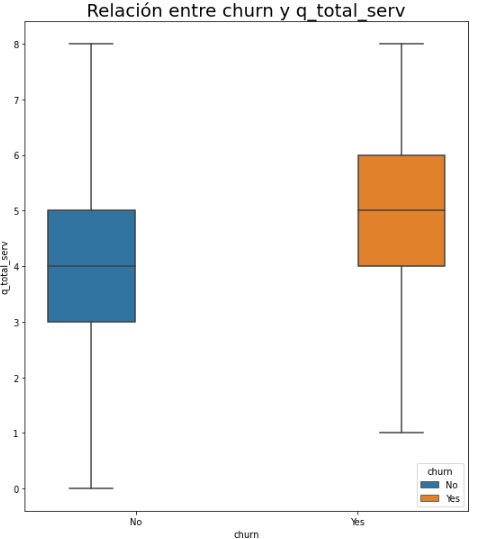
Podemos ver que quienes abandonan tienen una mediana más elevada y su rango intercuartílico también se ubica en valores mayores que quienes no abandonan, pero su dispersión es mayor, teniendo una distribución más uniforme. Su correlación es positiva (0.36). ¿Pueden ser personas que contraten más servicios o servicios más caros? Hay outliers con altos pagos mensuales en los que se quedan, es importante ver qué tipo de servicio estarán contratando estas personas.

**Churn Vs. Bandwidth GB Year**



A menor bandwidth hay más churn, puede ser que los servicios que brinda la empresa no funcionen bien con poco bandwidth o que los clientes no usen mucho los servicios. La correlación es de -0.39. ¿Las personas que usan menos Gb son mayores? ¿Por qué hay poca gente que usa 3000/3500?

**Churn Vs. Total Services**

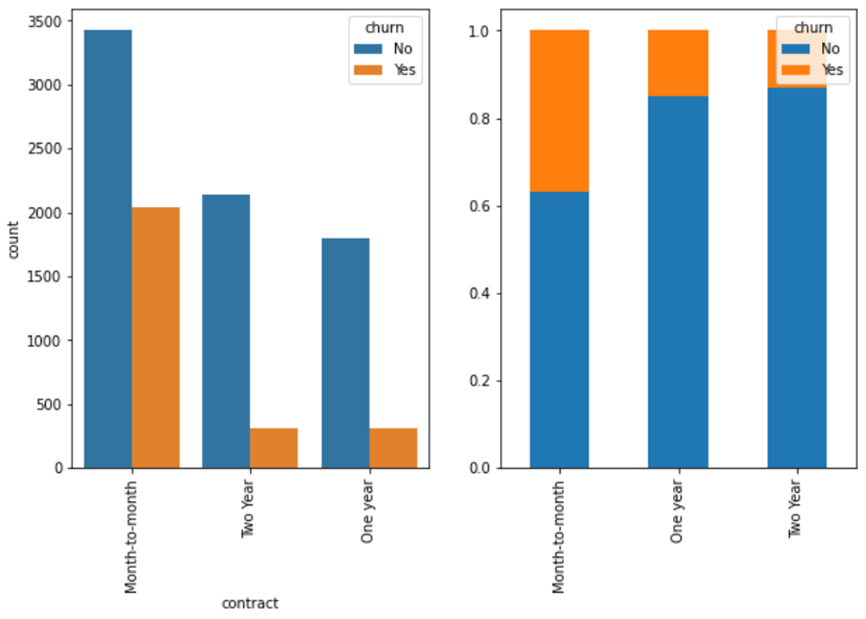


Quienes tienen 1 ó 2 servicios abandonan menos que quienes tienen más servicios ¿Por qué abandona alguien que tiene muchos servicios? ¿Les es fácil cambiarlos?

### Variable Churn Vs. Categóricas

A continuación, haremos mención de aquellas categorías en las que encontramos algún comportamiento diferente en cuanto al abandono (churn).

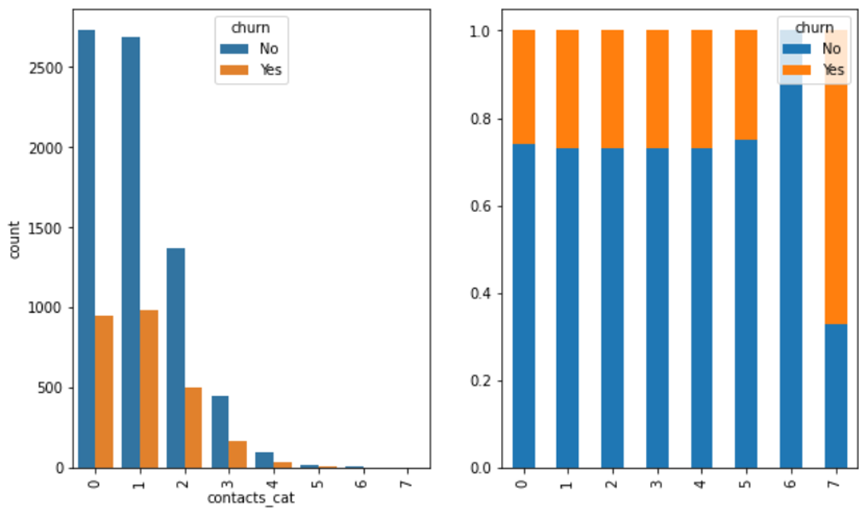
**Churn Vs. Contract**



Month-to-month contract presenta un porcentaje más elevado de churn, superando ampliamente a quienes contrataron servicios de uno y dos años. Esto puede deberse a que los contratos de uno y dos años son pagados por adelantado, por lo que la persona decidirá abandonar o no la empresa después de pasado el tiempo del contrato. Vale la pena agregar

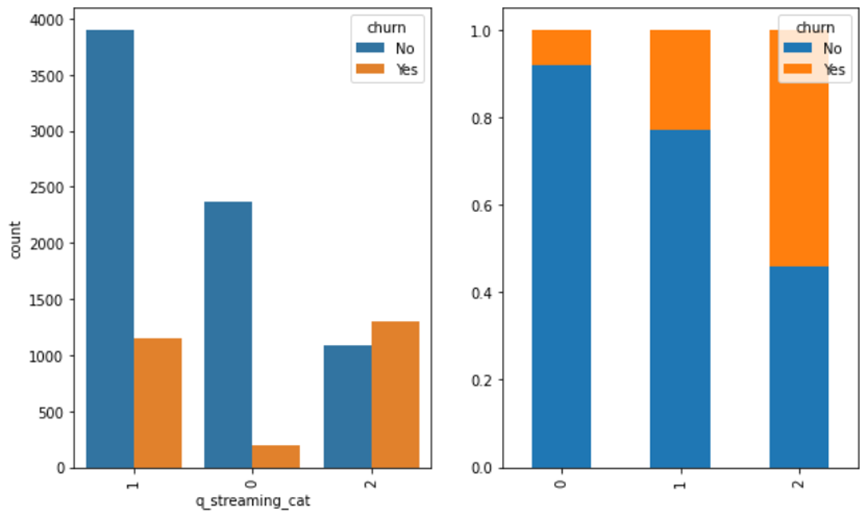
a esta comparación el tenure. Pregunta: ¿El abandono sigue siendo menor después de superado el tiempo del contrato? Ej. Alguien que contrató por un año, abandona después de 12 meses o por lo general decide continuar. ¿Varía según servicios que contratan?

**Churn Vs. Contacts**



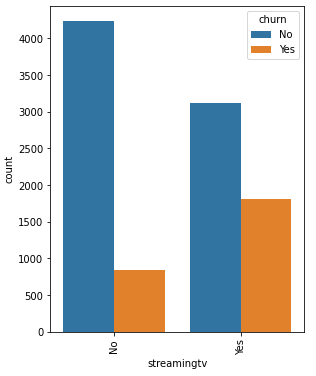
El 67% de quienes contactaron 7 veces a technical support, abandonaron. Si bien solo 3 clientes forman parte de la categoría, tiene sentido que si alguien se queja reiteradamente abandone. Más adelante podemos también agregar en el análisis la variable tenure, ya que no es lo mismo tener 3 quejas en 40 meses que 3 quejas en 3 meses. Esto cambia la percepción de la persona sobre el servicio.

**Churn Vs. Streaming**



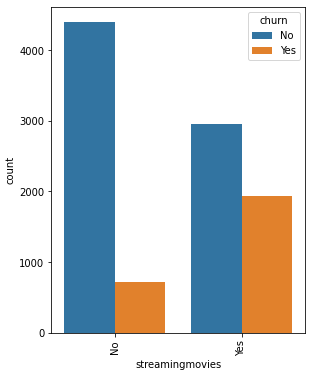
Las personas que contratan los dos servicios de streaming de la empresa, tienen una tasa de abandono del 54%.

**Churn Vs. Streaming Tv**



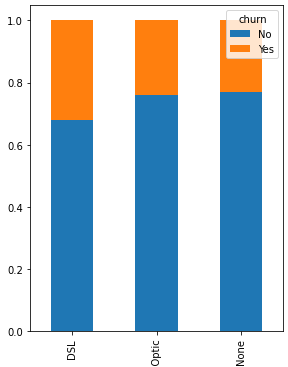
Dada la alta tasa de abandono según servicio de streaming, decidimos profundizar en cada uno de ellos. Respecto a streaming tv, la tasa de abandono en quienes si contratan el servicio es del 37% mientras que aquellos que no contratan, tiene un abandono del 16%. Aproximadamente la mitad de nuestros clientes contrata el servicio y la otra mitad no. ¿Será que encarece mucho la tarifa y por eso abandonan? ¿Qué tan buena será la calidad del servicio?

**Churn Vs. Streaming Movies**



En el caso de contratación de streaming movie, la tasa de abandono incrementa respecto a tv: quienes si contratan tiene una tasa del 40% mientras que los clientes que no lo contratan tienen una tasa del 14%. Aquí también se puede observar que es aproximadamente la misma cantidad de gente la que contrata el servicio que la que no contrata. ¿Será muy costoso este servicio? ¿Estará incluido en el abono del primer año y luego el cliente debe pagar y abandonar la empresa?¿Hay relación entre streaming movie y contratos anuales/bianuales?

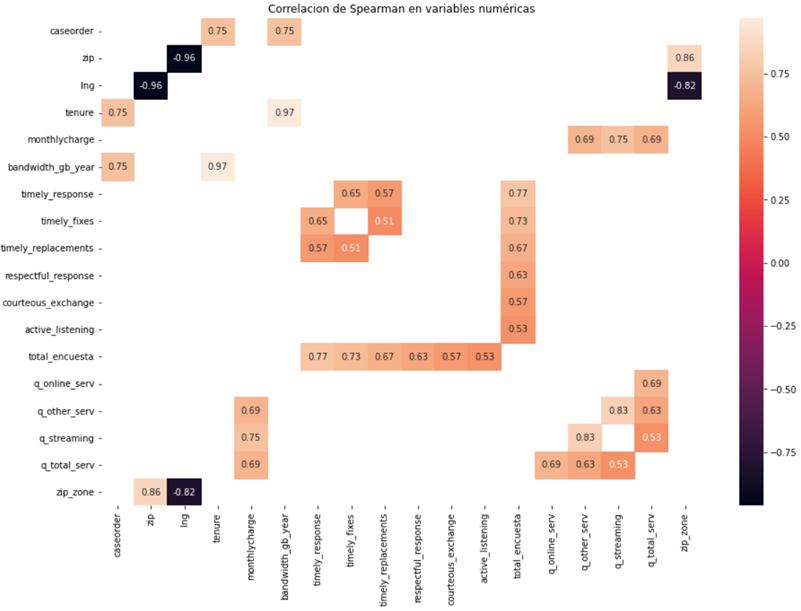
**Churn Vs. Internet Service**



El tipo de conexión a internet afecta en la churn siendo la tasa de abandono de un 32% en las personas que contratan por DLS mientras que internet por Fibra Óptica tiene una tasa de abandono del 24%. Quienes no contratan internet, tienen una tasa de abandono del 25%. ¿Será que no es bueno el servicio de DLS? ¿Las personas que contratan internet por DLS son las mismas que abandonan teniendo contratado streaming movie o streaming tv y abandonan por no poder ver televisión con buena calidad?

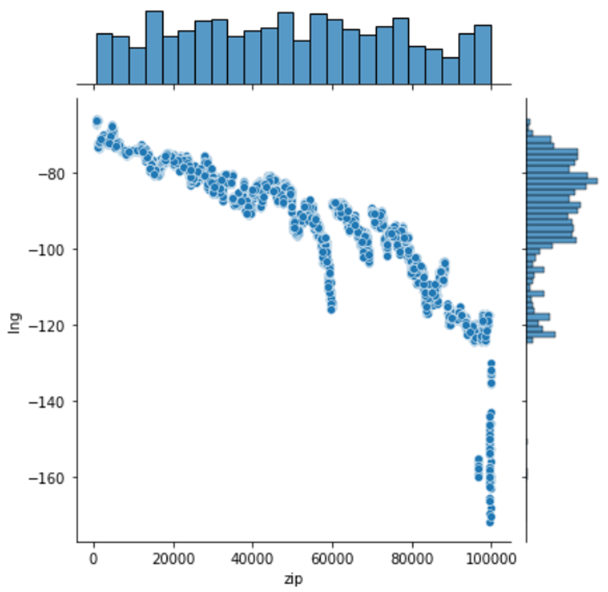
### Numéricas Vs. Numéricas

Armamos el Heatmap con la correlación de Spearman de las variables con relaciones +/- 0.5



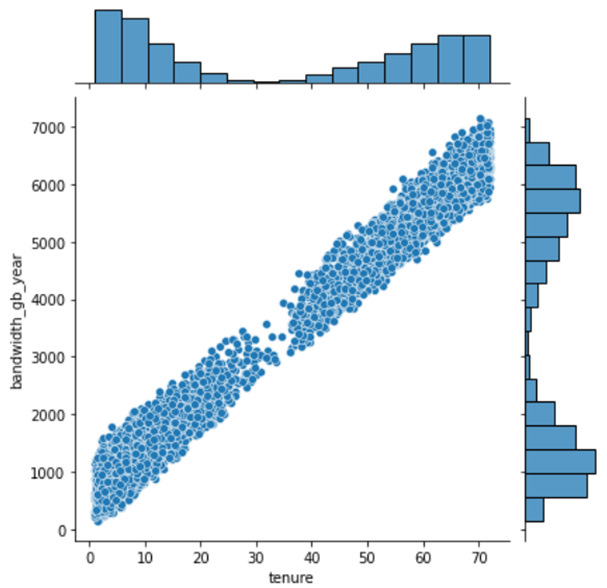
Luego armamos el joinplot para analizar estas variables, en términos generales no hay correlaciones significativas entre variables numéricas. Muchas de las encontradas se deben a que algunas variables se armaron custom a partir de otras variables. Ej: valores de encuestas y valor total de la encuesta.

**Zip Vs. Longitud**



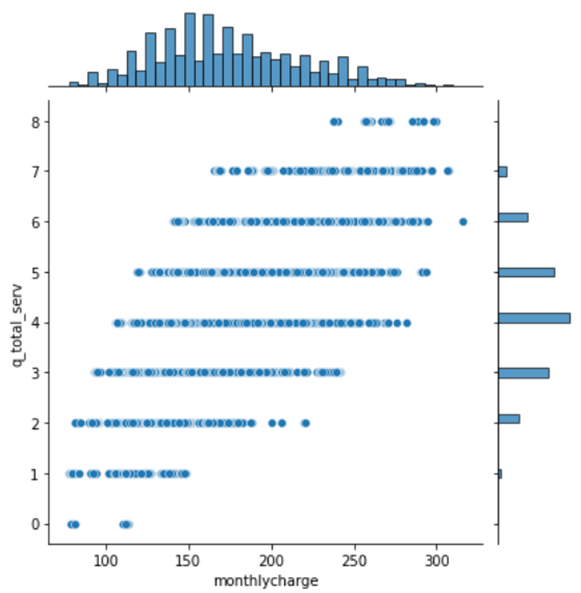
La correlación puede estar asociada a cómo se determinan los zip codes en EEUU.

**Bandwidth GB Year Vs. Tenure**



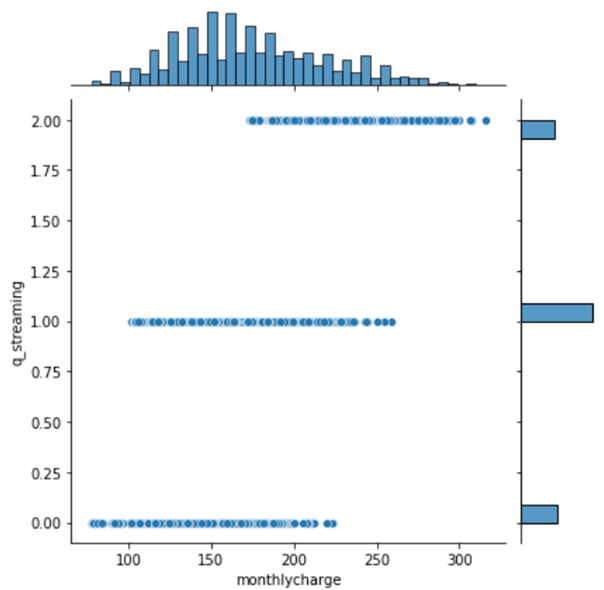
Tienen una correlación fuerte, ¿a qué se podrá deber? ¿Hay un límite de bandwidth mensual tope?

**Total Services Vs. Monthly Charge**



El monto total deriva de la cantidad de servicios contratados.

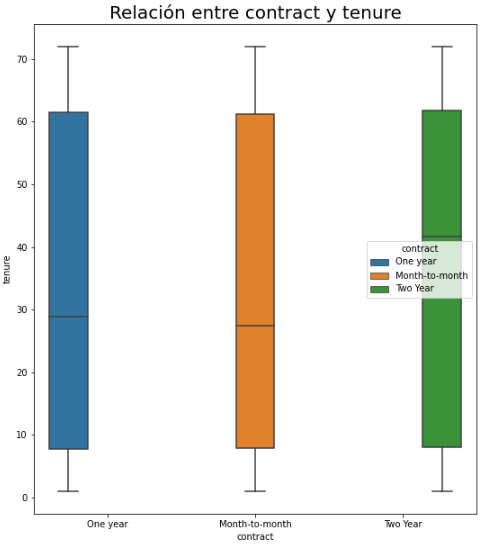
**Streaming Vs. Monthly Charge**



Tienen una correlación del 75%, esto indica que es significativo el costo de estos servicios por sobre el costo de los otros servicios ofrecidos por la empresa.

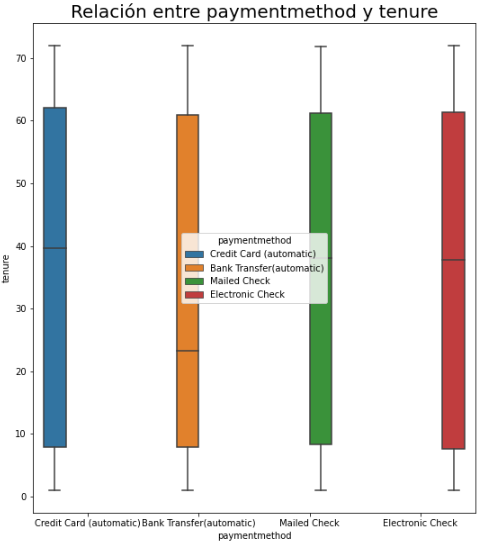
### Categóricas Vs. Numéricas

**Tenure Vs. Contract**



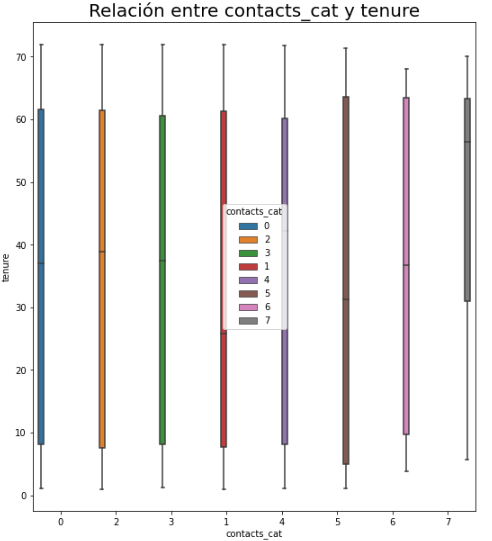
Month to month y one year tienen una mediana parecida, alrededor de 30. Quienes tiene contrato bianual, tienen una mediana de tenure mayor.

**Tenure Vs. Payment Method**



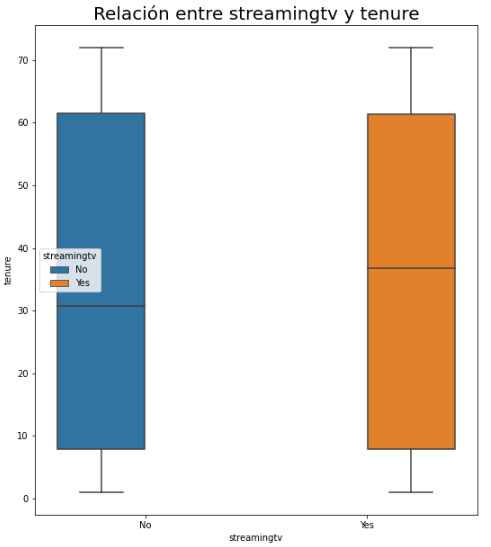
Quienes pagan con transferencia bancaria tienen una mediana más chica de antigüedad en la empresa.

**Tenure Vs. Contacts**



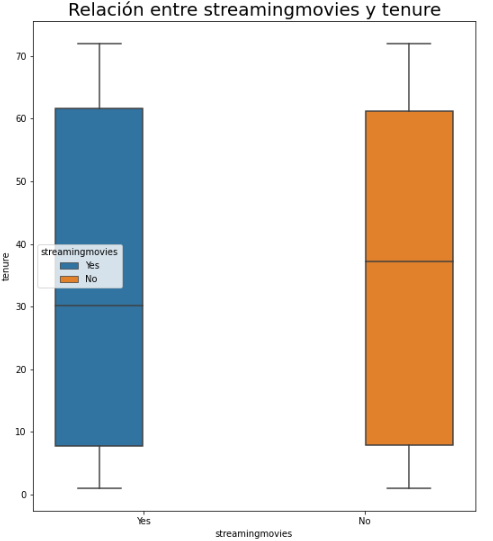
Quienes se contactaron 7 veces tienen mayor tenure.

**Tenure Vs. Streaming TV**



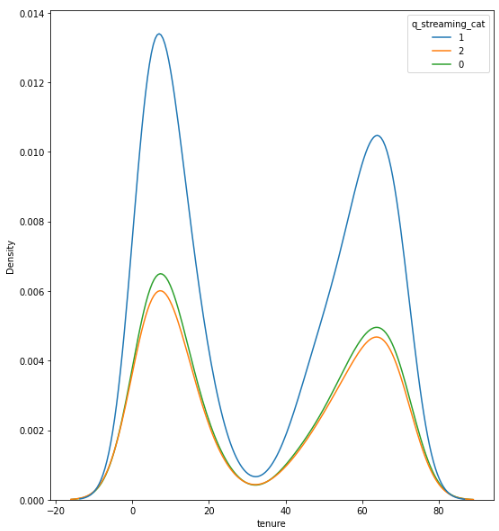
Quienes contratan streaming TV tienen un tenure mayor que quienes no contratan.

**Tenure Vs. Streaming Movies**



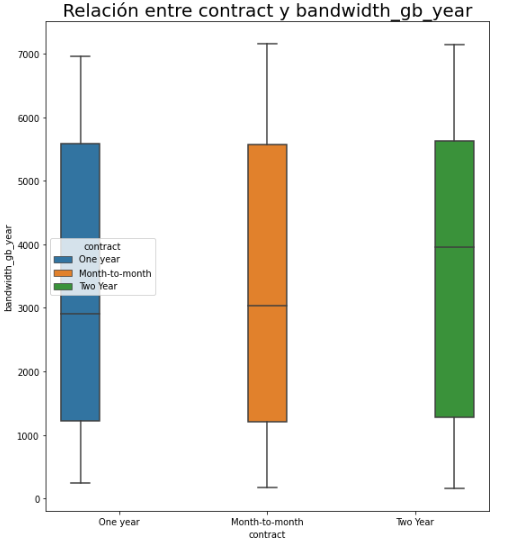
Quienes contratan streaming movies tienen un tenure menor que quienes no contratan este servicio.

**Tenure Vs. Streaming**



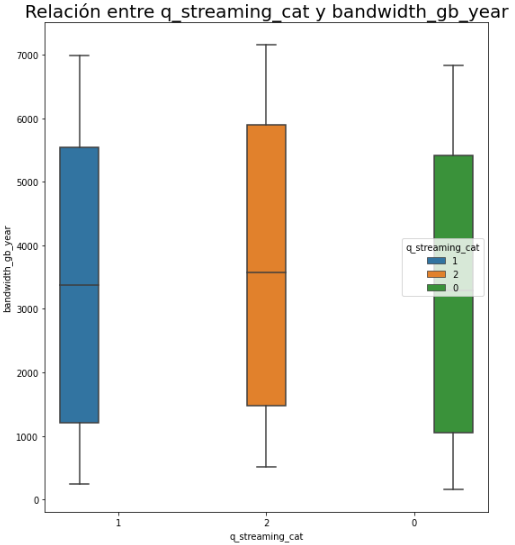
Quienes contratan un servicio de streaming tienen un tenure mayor que quienes contratan 2 servicios o ninguno. Es decir, tenemos dos tipos de clientes según consumo de streaming apenas contratan con la empresa: los que contratan y prueban todos los servicios y los que no contratan ninguno. ¿Será que aquellos que no contratan ninguno tiene más probabilidades de permanecer en la empresa que quienes contratan los dos servicios de streaming, siendo estos últimos personas más inquietas?

**Bandwidth GB Year Vs. Contract**



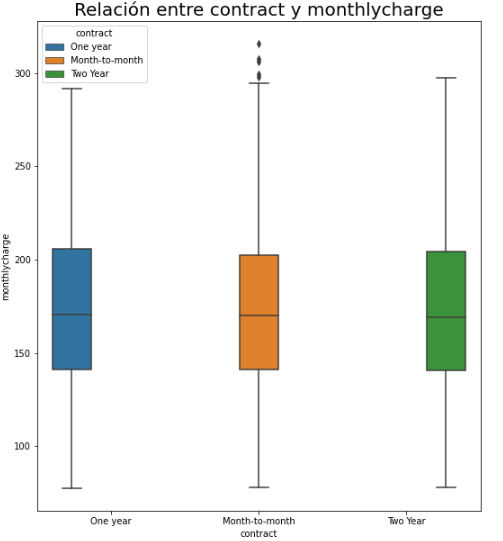
Quienes contratan por dos años tienen una mayor mediana.

**Bandwidth GB Year Vs. Streaming**



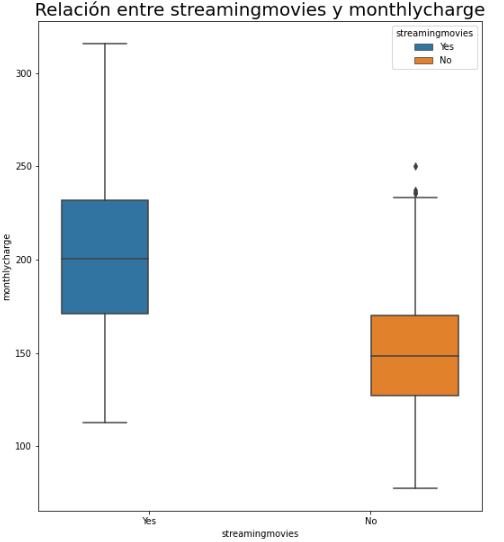
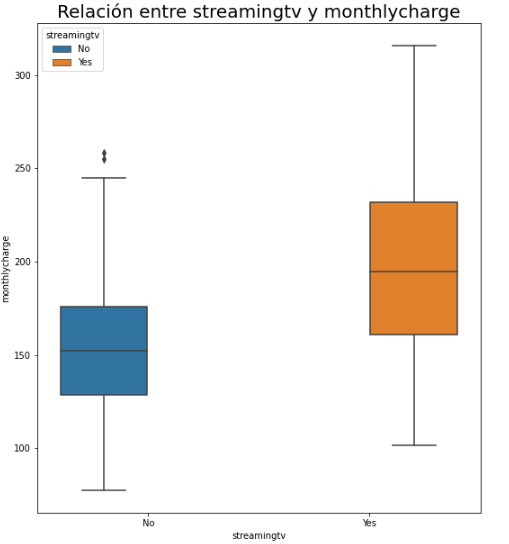
Quienes contratan streaming tienen mayor mediana, lo cual es esperado porque son servicios relacionados a videos con alto tráfico de datos.

**Monthly Charge Vs. Contract**



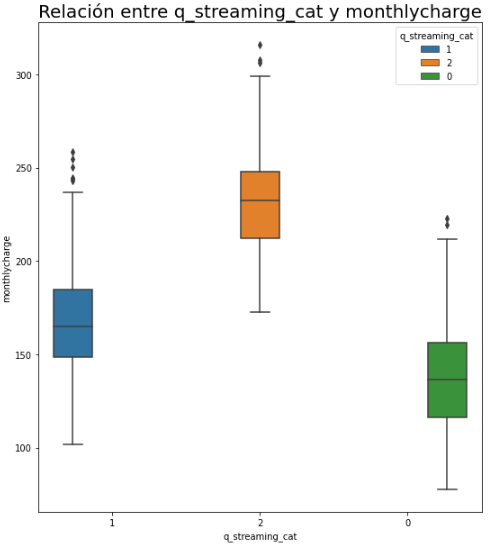
La mediana en los 3 tipos de contrato es similar. Esto indica que el cliente no paga menos por tener un contrato anual o bianual. ¿Cuál será la ventaja de tener un contrato de mayor duración?

**Monthly Charge Vs. Streaming TV/Movies**



La mediana del abono es similar para quienes contratan streaming TV y Movies. La mediana en ambos casos es de USD 200, mientras que quienes no contratan alguno de estos servicios tienen una mediana de USD 150.

**Monthly Charge Vs. Streaming**

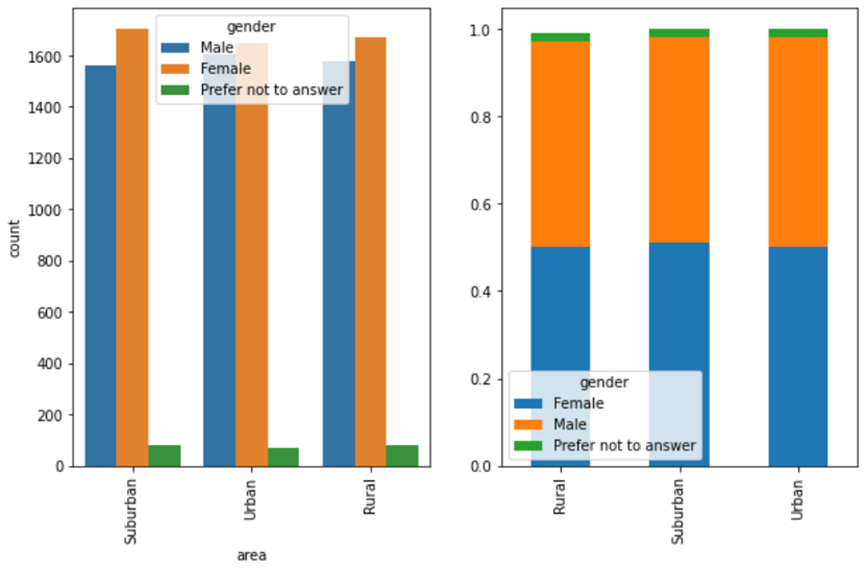


Quienes contratan los 2 servicios de streaming tienen una mediana mensual de USD 240.

### Categóricas Vs. Categóricas

El análisis entre este tipo de variables no nos estaría sumando en términos de calidad, son más bien triviales comparados con lo que ya hemos visto. De todos modos, colocamos el siguiente ejemplo.

**Area Vs. Género**



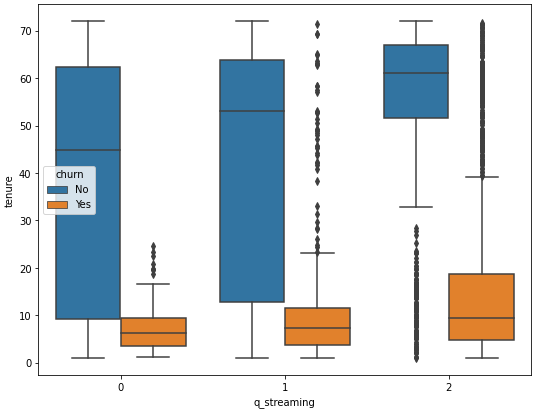
Las proporciones por género se mantienen según sea el área donde se brinde el servicio.

## Análisis Multivariado

En esta etapa del EDA se analizan 3 ó más variables al mismo tiempo para comprender en profundidad su comportamiento.

Buscaremos responder las preguntas que surgieron tanto en el análisis univariado como en el bivariado siguiendo de esta manera las pistas que nos va brindando la exploración del dataset.

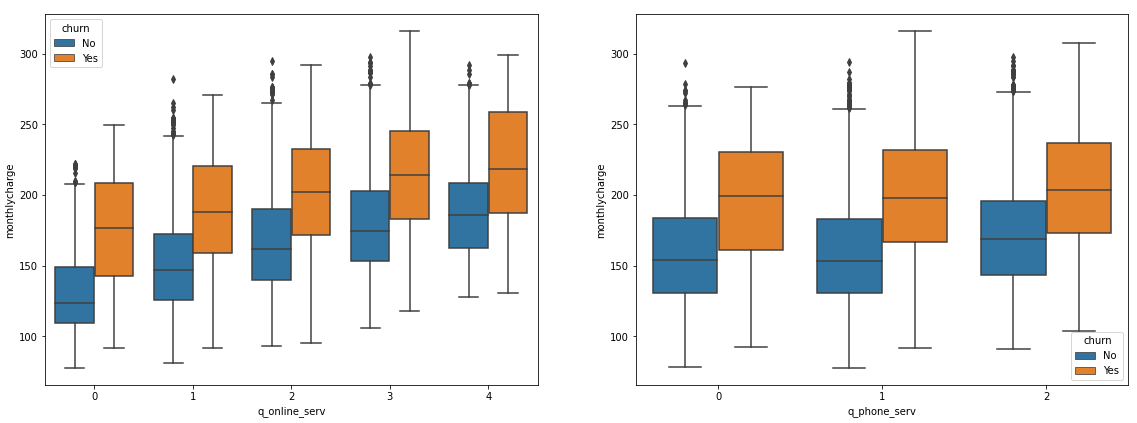
**Streaming, tenure y churn:**



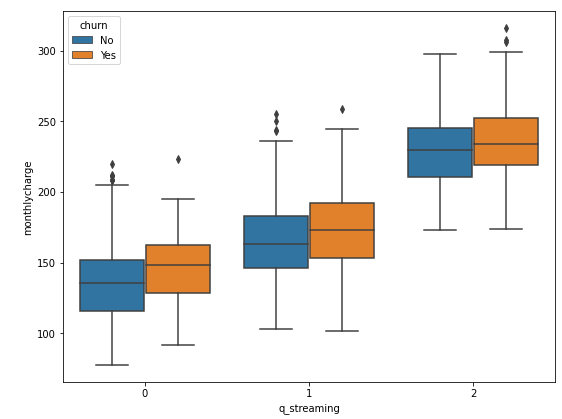
Es notable la concentración de datos en los casos que contratan dos servicios de streaming: se concentran en niveles bajos de tenure y en niveles altos de tenure. Aquellos que tienen muchos meses de antigüedad no abandonan al contratar streaming pero si lo hacen quienes tienen pocos meses de antigüedad. Una posibilidad es que se encarezca mucho el servicio y, al no estar fidelizados por antigüedad, deciden cambiar de empresa.

Se observan también muchos outliers: clientes con tenure por encima de los veinte meses y a pesar de la antigüedad, contratan uno o dos servicios de streaming y abandonan la empresa. En los casos de no churn, se observan muchos outliers en quienes contratan dos servicios de streaming y tienen pocos meses de antigüedad. ¿Por qué se quedan esas personas?

**Online services, phone service vs monthly charge y churn:**



**Streaming vs monthly charge y churn:**

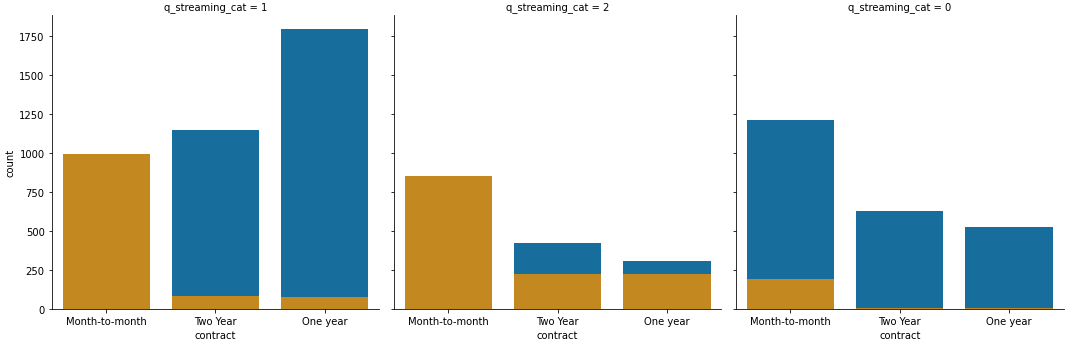


Del análisis bivariado surge la pregunta de qué tipo de servicio contratan los outliers que pagan mucho y no abandonan. Al observar estos gráficos podemos afirmar que tenemos outliers con pagos mensuales altos en todos los servicios. Solo no hay outliers en quienes consumen los dos servicios de streaming.

Al comparar estos tres gráficos que analizan como variable dependiente la categoría monthly\_charge se observa una disparidad en la mediana: quienes abandonan tienen mediana considerablemente mayor en el abono.La diferencia en el abono en usd entre quienes abandonan y los que no es de 50usd aprox y se mantiene para los distintos tipos y cantidad de servicios contratados. Esta mediana superior en todos los casos confirma que para quienes contratan estos servicios si es determinante en aumento en el abono para abandonar.

Este gráfico nos permite observar que la diferencia en el costo del abono entre los que dejan y los que no dejan es de 10usd aprox en el caso de los servicios de streaming. Esto indicaría que en el caso de los servicios de streaming no es el costo del abono lo que hace que abandonen la empresa.

**Servicio de streaming, contract y churn:**



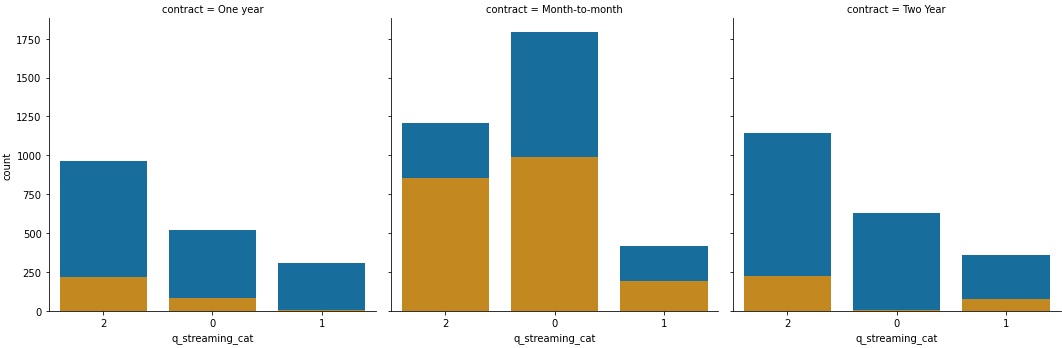
De los clientes que tienen contrato ‘month to month’ solo continúan en la empresa quienes no contratan streaming. Quienes contratan uno o dos servicios de streaming y tienen contrato mensual, abandonan la empresa. Será que lo tiene bonificado los primeros meses y luego deben pagarlo y se dan de baja de la empresa?

Quienes contratan un servicio de streaming y tienen contrato anual o bianual, no abandonan la empresa en general. Es notable la fidelización de aquello que tiene contrato anual y contratan un servicio de streaming. ¿Cuál será la antigüedad de estos clientes?

Otro dato relevante es la baja cantidad de churn en aquellos que no tienen contratado servicio de streaming.

Quienes contratan dos servicios de streaming, independientemente del tipo de contrato, son los que más abandonan.

.

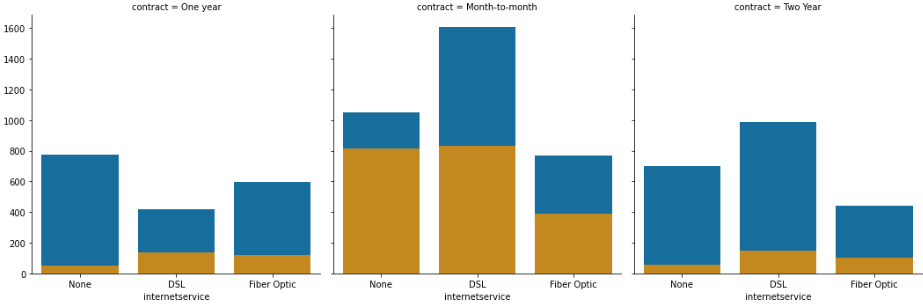


En este gráfico hemos alternado la posición de contract y streaming lo cual nos permite observar con facilidad la concentración de abandono en quienes tienen contrato mensual.

También se observa un 75% aprox de abandono en quienes tiene contrato mensual y contratan los dos servicios de streaming disponibles. En los otros dos tipos de contrato es baja la tasa de abandono. Si se observa que es mayor en quienes contratan ambos servicios de streaming.

Quienes contratan por dos años y no tiene sv de streaming tiene muy baja churn. Casi el total de la muestra continua en la empresa. Lo mismo para quienes tienen contrato anual y contratan un solo sv de streaming.Tal como vimos en el análisis bivariado, quienes tiene contrato anual y bianual abandonan menos la empresa.

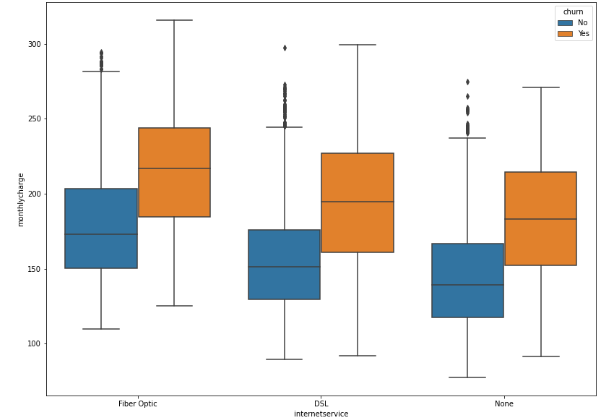
**Servicio de internet, contract y churn:**



Este gráfico analiza el comportamiento de los clientes respecto al tipo de servicio de internet contratado. Resaltamos la mayor cantidad de clientes con contrato mensual por sobre los otros contratos. Estas personas contratan en su mayoría DSL service, de ellos, la mitad abandona. Le siguen en cantidad de personas quienes no contratan internet. De estas personas, el 75%abandona la compañía. Son menos las personas con contrato mensual que contratan fibra óptica. De ellos, la mitad abandona.

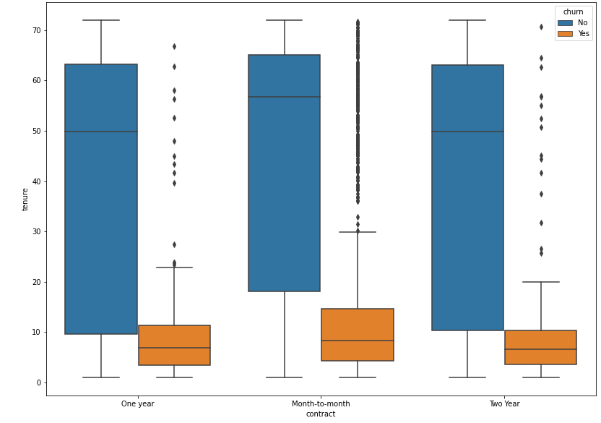
Los contratos anuales y bianules tiene muy baja tasa de abandono independientemente del tipo de internet.

**Servicio de internet, monthly charge y churn:**



Internet por fibra óptica tiene una mediana de abano superior a DSL. Quienes abandonan la empresa tienen en todos los casos una mediana de abono superior que quienes se quedan.

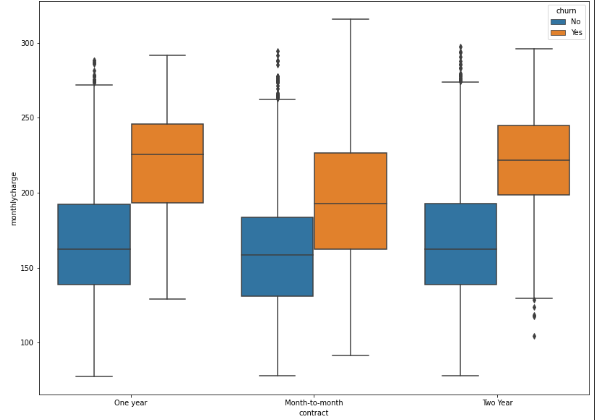
**Contract, tenure y churn:**



En este gráfico se observa una mediana de tenure mayor cuando el contrato es mensual tanto en quienes abandonan las empresas como los que se quedan. El hecho de tener un contrato por plazo más corto no indica que vayan a dejar antes la empresa. ¿Por qué no establecen entonces contratos a más largo plazo? ¿no obtiene beneficios por contratos más largos?

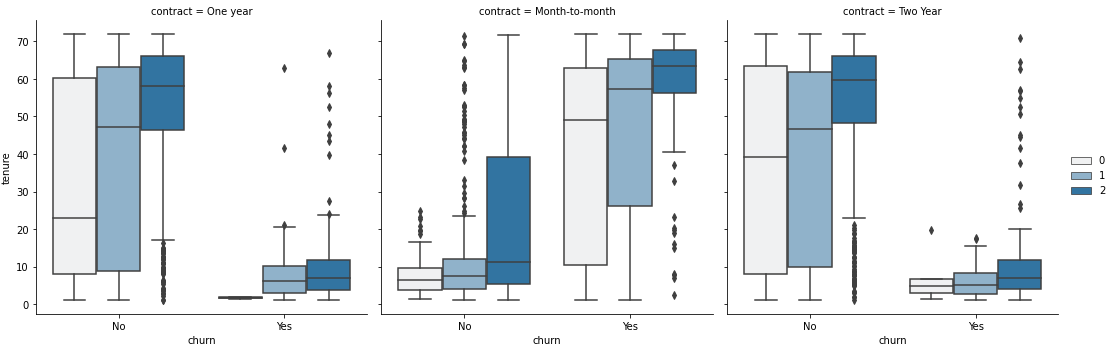
Llama la atención de este gráfico la concentración de los datos en tenures bajas en todos los casos de churn independientemente del tipo de contrato.

**Contract, monthly charge y churn**



Este gráfico demuestra que quienes abandonan la empresa siempre tienen medianas de abono mayor que quienes continúan. Observamos también que quienes tienen contrato mensual tienen una mediana menor que quienes tienen contrato anual y bianual. Por tanto, pagan menos. Será que al empezar el contrato se les bonifica y al cabo de unos meses deben pagar más y por eso se van?

**Servicios de Streaming, tenure, contract y churn**



Este gráfico nos permite observar el comportamiento de antigüedad en las personas que abandonan la empresa respecto al tipo de contrato que tienen y la contratación de streaming. Observamos que quienes tienen contrato mensual abandonan en tenures mucho más altas que quienes tienen contrato anual y bianual independientemente de si contratan streaming o no. ¿Que lleva a esta gente a abandonar si no es el servicio de streaming? Si se observa que a medida que incrementa la cantidad de sv, incrementa la mediana de la tenure.

Los abandonos en contratos mensuales tienen una mediana de antigüedad de 60 meses mientras que en contratos anuales y bianuales los abandonos se producen con una mediana de 9 meses aprox. Abandona antes que termine el contrato. No parece verse muy afectado por la cantidad de sv que contratan.

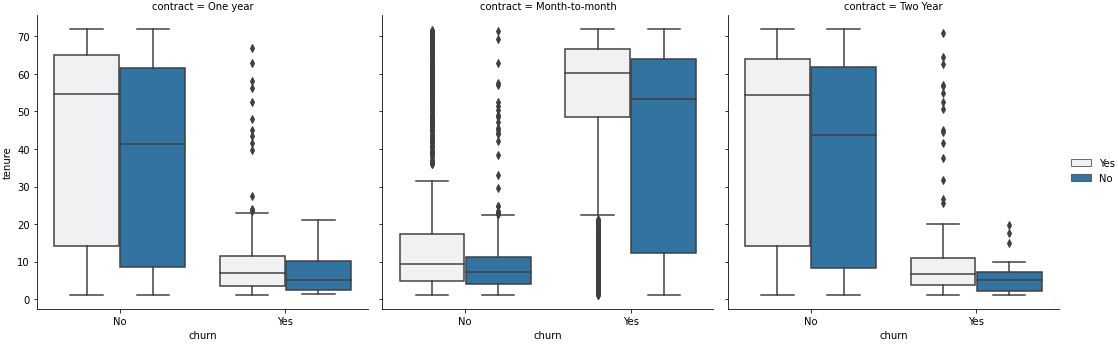
Tanto en anual como bi anual, las personas que no abandonan tienen una tenure alta.

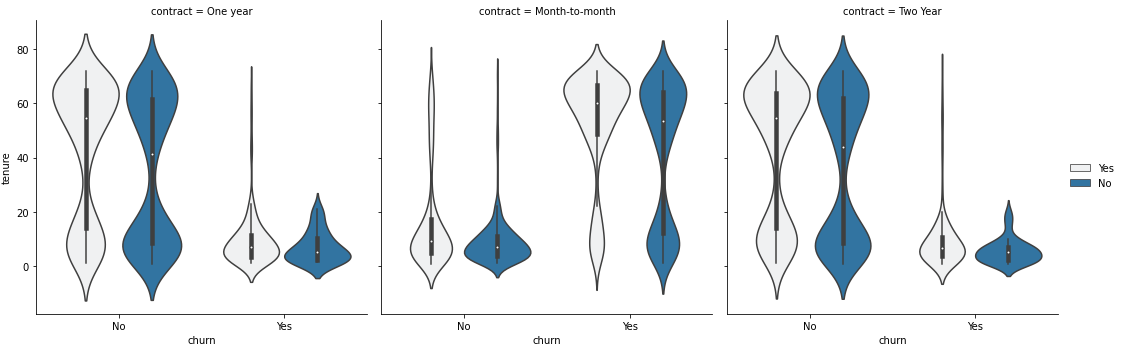
En cambio, en los contratos mensuales que no abandonan, la mediana de la tenure es considerablemente baja. Estas personas, cuando alcancen tenures más altas, serán nuestros próximos churn?

Cabe destacar el comportamiento de quienes tienen contrato por dos años y contratan muchos servicios, abandonan con tenures más altos que quienes nos abandonan con menos servicios. Quizá sí sea una variable de retención la cantidad de sv contratados en este caso.

En los contratos anuales y bi anuales se observan datos muy dispersos respecto a la tenure en aquellos que no abandonan. Quienes si abandonan se concentran en tenures bajas, no hay dispersión de datos.

**Streaming movies, tenure, contract y churn**





Sabiendo que el 40% de las personas que contratan streaming movies abandonan la empresa mientras que quienes no contratan abandonan solo el 14%, hemos decidido profundizar en esta variable.

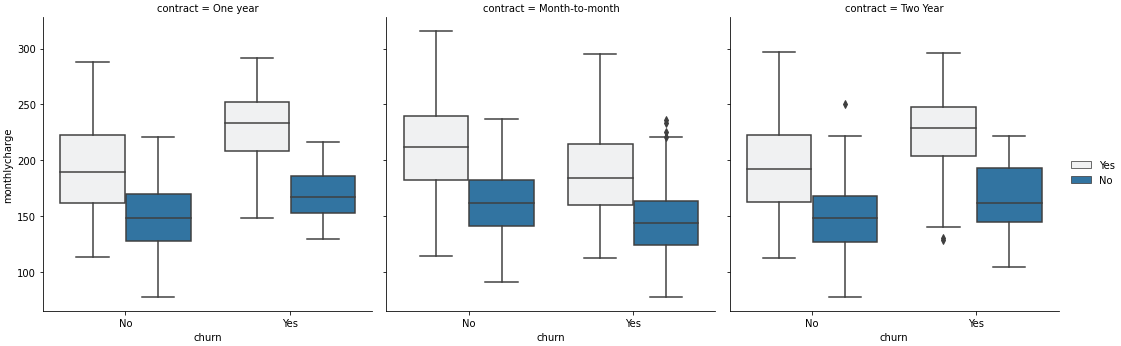
Quienes si contratan y nos abandonan tienen en su mayoría una contrato mensual como se observa en la curvatura del violín plot. Estas personas nos abandonan con tenure alta. Existe un porcentaje de personas, consideradas outliers por el boxplot, que contratan mensual y abandona con tenures bajas.

En los contratos anuales y bi anuales, la mayor cantidad de personas que abandonan son las que no contrataron servicio de streaming movie. Y menor cantidad de churn para quienes sí contrataron. Ambos grupos abandonan generalmente antes de llegar al año de antigüedad.

Se observa gran dispersión de datos en las personas con contrato mensual que abandonan la empresa sin haber contratado servicio de streaming: tenemos personas que se van con tenures bajas y personas que se van con tenures altas.

En todos los tipos de contrato tanto como si son churn o no, quienes contratan servicio de streaming movies tienen una mediana de tenure mayor.

**Streaming movies, monthly charge, contract y churn**



Respecto al abono mensual,los que contratan streaming movie siempre tienen medianas mayores.

En los contratos anuales y bianuales se cumple que abandonan con medianas de abono más altas que quienes no abandonan. El abono parecería ser influyente en estos tipos de contrato.

En los contratos mensuales, el abandono se produce en medianas más bajas que quienes permanecen. Por tanto, quienes no son churn, tienen un abono mensual mayor que los que abandonan, no parecía afectarles el pago en el caso de contratos mensuales.

## Conclusion

Las variables que más afectan a la decisión de abandonar o no la empresa parecerían ser:

* Tenure: la mayor tasa de abandono se da en tenures bajas, con una mediana de 10 meses.
* Monthly charge: Quienes abandonan la empresa tienen una mediana de abono mensual mayor que quienes no abandonan. Solo para el caso de los contratos mensuales, el abandono se produce en medianas más bajas que quienes permanecen. Por tanto, quienes no abandonan, tienen un abono mensual mayor que los que abandonan, no parecía afectarles el pago en el caso de contratos mensuales.
* Contract: la mayoría de nuestros clientes tiene contrato mensual. Este tipo de contrato tiene una tasa de abandono del 40% aprox mientras que los contratos anuales y bi anuales tienen una tasa de abandono del 15%.Los abandonos en contratos mensuales tienen una mediana de antigüedad de 60 meses mientras que en contratos anuales y bianuales los abandonos se producen con una mediana de 9 meses aprox. Abandona antes que termine el contrato.
* Streaming Movie: Las personas que contratan este servicio tienen una tasa de abandono del 40%. Estas personas tienen en su mayoría un contrato mensual y nos abandonan con tenure alta. En los contratos anuales y bi anuales, la mayor cantidad de personas que abandonan son las que no contrataron servicio de streaming movie.
* Internet Service: quienes contratan DLS service tiene una tasa de abandono 10% mayor que quienes contratan fibra óptica o quienes no contratan internet. El servicio de DLS es más barato que la fibra óptica.

# Desarrollo del Modelo

Tal como describimos en el objetivo del trabajo, el mayor desafío que enfrenta la empresa es lograr retener sus clientes actuales reduciendo al mínimo posible su tasa de abandono(variable ‘churn’). Para lograr esto, debe identificar aquellos clientes con mayor probabilidad de churn para poder ofrecerles una propuesta de servicio superadora que los retenga. Es por ello que nuestro problema es del tipo supervisado y se resuelve con modelos de clasificación.

Algoritmos utilizados:

* Random Forest
* Regresion Logistica

La métrica que nos interesa sea lo mejor posible es el **f-score** pues se trata de un problema desbalanceado. Buscamos también un alto **recall** ya que queremos predecir la mayor cantidad de personas que nos van a abandonar siendo menos importante en nuestro problema la precisión.

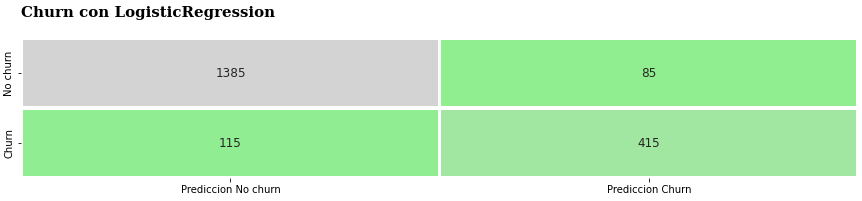
## Random Forest

Matriz de confusion:



## Regresion Logistica

Matriz de confusion:



## Metricas

| Algoritmo | Precision | Recall | f1-score | auc |
| --- | --- | --- | --- | --- |
| Random Forest | 0.84 | 0.74 | 0.79 | 0.84 |
| Regression Logistica | 0.83 | 0.78 | 0.81 | 0.86 |

Debido a la importancia que tiene para este problema de negocio el recall y el f-score, seleccionamos el algoritmo de Regresión Logística