UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas /UPV

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas

El triunfo de los vídeos cortos en RRSS. Factores que inciden en las primeras recomendaciones de TikTok para cuentas recién registradas.

Trabajo Fin de Máster

Máster Universitario en Social Media y Comunicación Corporativa

AUTOR/A: Garcia Fernandez, Kevin

Tutor/a: Crespo Abril, Fortunato

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MÁSTER EN: SOCIAL MEDIA Y COMUNICACIÓN CORPORATIVA

2023/2024

TRABAJO FIN DE MÁSTER

EL TRIUNFO DE LOS VÍDEOS CORTOS EN RRSS. FACTORES QUE INCIDEN EN LAS PRIMERAS RECOMENDACIONES DE TIKTOK PARA CUENTAS RECIÉN REGISTRADAS.

AUTOR: KEVIN GARCÍA FERNÁNDEZ

TUTOR: FORTUNATO CRESPO ABRIL

VALENCIA, 29 DE NOVIEMBRE DE 2023

ADE
FACULTAD DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS. UPV

RESUMEN

Este trabajo intenta abordar los factores que inciden en las primeras recomendaciones de vídeos en TikTok cuando se trata de cuentas recién registradas. TikTok es una red social que basa su forma de mostrar contenido en una IA que recaba información de la interacción de la cuenta, es por esto por lo que surge la incertidumbre de qué ocurre cuando la cuenta es totalmente nueva y la aplicación no ha podido todavía recabar información sobre ella.

El estudio tiene como objetivo analizar qué factores tienen relevancia a la hora de que la plataforma muestre un vídeo a cuentas que no han tenido ninguna interacción previa. Además, también se busca diferenciar características propias de los vídeos y de los perfiles que se muestran, con el fin de formar una visión general del tipo de contenido al que se expone a estas cuentas.

Previo al análisis, se presenta la revisión bibliográfica sobre el auge que han tenido los vídeos cortos en redes sociales a raíz del triunfo de TikTok. Esta plataforma ha marcado un antes y un después en el mundo de las redes sociales, haciendo que el contenido en formato vídeo se imponga frente a las imágenes. La gran acogida de esta plataforma por parte de los usuarios se debe a su innovadora forma de mostrar el contenido a través de un algoritmo que utiliza inteligencia artificial, siendo una gran incógnita cómo funciona.

Para realizar el estudio se ha recabado información de las diez primeras recomendaciones realizadas por el algoritmo a varios perfiles recién registrados, teniendo en cuenta tanto los vídeos recomendados como los perfiles a los que pertenecen estos vídeos. Hay que tener en cuenta las limitaciones del estudio ya que factores como la fecha de creación de los perfiles, el área geográfica y el dispositivo desde el que se recoge información influyen en los resultados.

El análisis revela que la temática por excelencia de los vídeos es el humor y , principalmente, la persona que protagoniza los vídeos es una mujer. El rango de edad de los perfiles que se muestran con mayor frecuencia se ubica en mayores de 25 años, o entre los 18 y los 25 años, situándolos, por tanto, en la generación Z. Por otra parte, no necesariamente se muestran vídeos de perfiles con gran relevancia ya que la mayoría de ellos no disponía de la verificación de TikTok, conocida como *check azul*.

El estudio revela que existen diferencias significativas entre algunas de las variables recogidas, lo que nos permite formar una pequeña visión de cómo funciona el algoritmo de esta red social. Entre ellas, lo más destacado es que la relación entre las visualizaciones y el número de seguidores es relativamente débil, lo que apoya la teoría de que en TikTok no hace falta tener seguidores para obtener un gran alcance. También que el sexo del perfil influye en los comentarios de los vídeos recomendados, ya que si el perfil pertenece a una mujer existe mayor predisposición a comentar.

Finalmente se han utilizado dos modelos predictivos: regresión logística y árboles de decisión, en los que se pretende predecir si la plataforma TikTok recomienda o no, varios vídeos de un mismo perfil, en función de otras variables, con el objetivo de comprender mejor el algoritmo de la plataforma.

En definitiva, con este estudio se intenta entender el funcionamiento del algoritmo de recomendación de vídeos utilizado por TikTok cuando la plataforma no dispone de información sobre el usuario al que tiene que enviar contenido. Además, se hace énfasis en las características de los vídeos que recomienda y los perfiles a los que pertenecen esos vídeos.

Palabras clave: redes sociales; vídeos cortos; TikTok.

ABSTRACT

This paper attempts to address the factors that affect the first video recommendations in TikTok when it comes to newly registered accounts. TikTok is a social network that bases its way of displaying content on an AI that collects information from the interaction of the account, which is why uncertainty arises about what happens when the account is brand new and the application has not yet been able to collect information about it.

The study aims to analyze what factors are relevant when the platform shows a video to accounts that have not had any previous interaction. In addition, it also seeks to differentiate the characteristics of the videos and profiles that are shown, in order to form an overview of the type of content to which these accounts are exposed.

Prior to the analysis, a literature review is presented on the boom of short videos in social networks following the triumph of TikTok. This platform has marked a before and after in the world of social networks, making video content prevail over images. The great reception of this platform by users is due to its innovative way of displaying content through an algorithm that uses artificial intelligence, being a great unknown how it works.

To carry out the study, information was gathered from the first ten recommendations made by the algorithm to several newly registered profiles, taking into account both the recommended videos and the profiles to which these videos belong. The limitations of the study must be taken into account since factors such as the date of creation of the profiles, the geographical area and the device from which information is collected influence the results.

The analysis reveals that the theme par excellence of the videos, as expected, is humor and mainly the person starring in the videos is a woman. The age range of the profiles shown is mainly over 25 years old or between 18 and 25 years old, which means that it can be placed in the Z generation. On the other hand, videos of profiles with high relevance are not necessarily shown as most of them did not have TikTok verification, known as blue check.

In addition, it reveals that there are significant differences between some of the variables collected, which allows us to form a small vision of how the algorithm of this social network works. Among them, the most notable is that the relationship between views and the number of followers is relatively small, which supports the theory that in TikTok it is not necessary to have followers to obtain a large reach. Also, the gender of the profile influences the comments on the recommended videos, since if the profile belongs to a woman there is a greater predisposition to comment.

Two predictive models have also been designed to determine whether the TikTok platform recommends the same profile several times from the recommended videos or not based on other variables, with the aim of better understanding the platform's algorithm.

In short, this study attempts to understand the functioning of the video recommendation algorithm used by TikTok when the platform does not have information about the user to whom it has to send content from its platform. In addition, emphasis is placed on the characteristics of the videos it recommends and the profiles to which those videos belong.

Keywords: social networks; short videos; TikTok.

ÍNDICE DE TÉRMINOS CLAVE

Challenges: este término hace referencia a todo tipo de retos que vayan surgiendo en redes sociales, principalmente suelen tratarse de bailen con fragmentos de alguna canción en tendencia. No obstante, no solo se tratan de bailes, abarca una gran variedad de tema, como puede ser contar historias que haya vivido el usuario o realizar algún tipo de prueba.

Fast content: es un contenido que es creado específicamente para ser de rápido consumo. Suele tratarse principalmente de vídeos que muestran un mensaje breve, muy concreto y que aporte valor al espectador. Además, suele ir relacionado con tendencias del momento y prácticamente efímeras.

Feed: este término hace referencia a la visión general de todo el contenido que un usuario en redes sociales tiene subido online.

Hashtags: son las etiquetas que se utilizan para categorizar contenido dentro de redes sociales, y se utilizan comenzando con el símbolo: #.

Likes: es el anglicismo de los conocidos *me gustas* en redes sociales, cuando una persona reacciona a cualquier contenido en redes sociales mostrando interés por él.

Social media marketing: este concepto se trata de todas las acciones y estrategias del marketing digital que están dirigidas y enfocadas a las redes sociales.

Views: anglicismo que se relaciona con las visualizaciones que pueda tener un contenido, en este estudio hace referencia a las visualizaciones de los vídeos de TikTok.

Check azul: se trata de una verificación de la propia plataforma que otorga a perfiles de personas con gran repercusión o a marcas con el objetivo de verificar que es la cuenta oficial.

Influencer: en el contexto de este trabajo hace referencia a una persona con repercusión en redes sociales con la capacidad de influir en usuarios, principalmente en sus seguidores.

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES E IMÁGENES

1. INTRODUCCIÓN

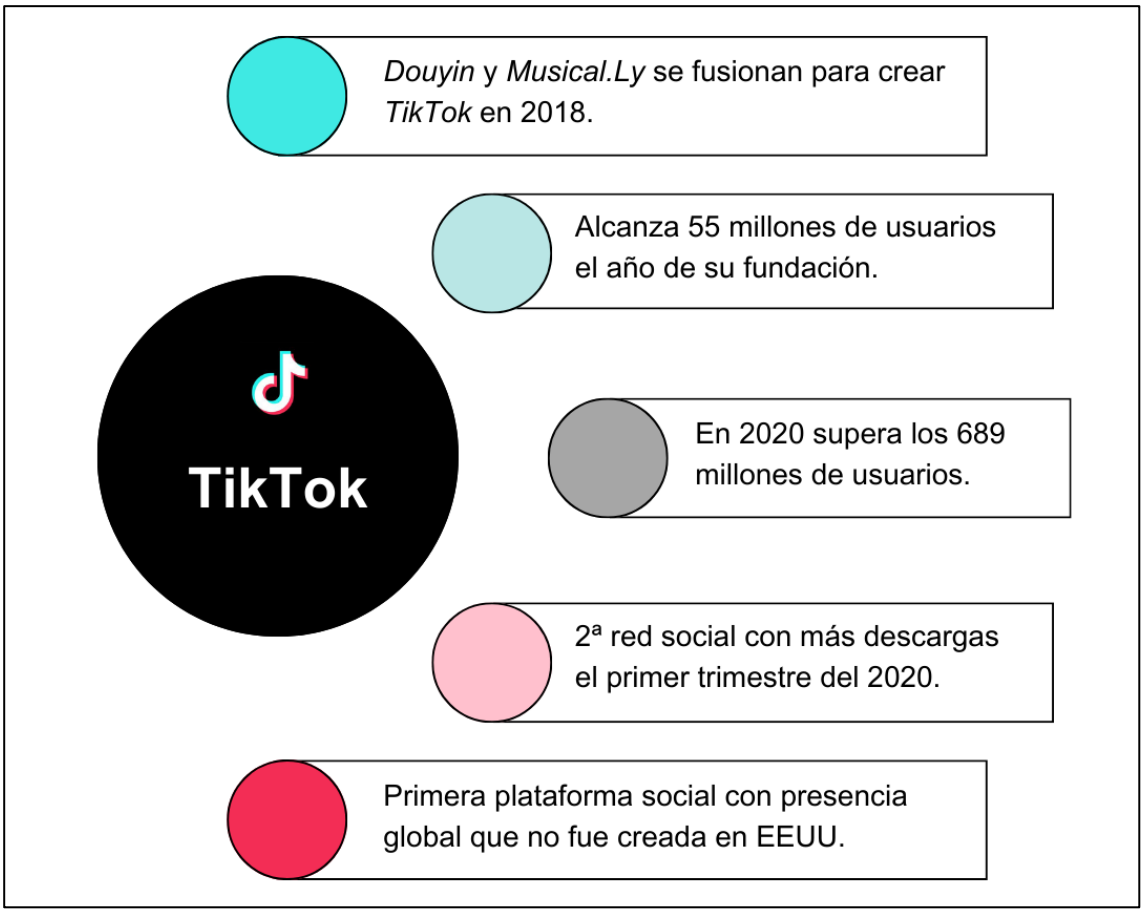
1.1. TIKTOK: SU HISTORIA

Actualmente, no se puede hablar del *social media marketing* sin tener en cuenta la red social favorita de la generación Z, TikTok. La gran evolución que ha tenido esta red social y cómo ha logrado posicionarse como una de las plataformas principales, es un hito que ha marcado tendencia en el mundo de las redes sociales.

TikTok es una red social basada en vídeos verticales cortos y de uso exclusivo en aplicación para móvil. Sus funciones principales son la edición y difusión de dichos vídeos que pueden llegar a un gran número de visualizaciones gracias al algoritmo que utiliza (Martin, Merino y Micaletto, 2022; Méndez y Olivares, 2020).

El origen de esta red social surge en 2018 cuando las plataformas *Douyin* y *Musical.Ly* deciden fusionarse. El año de su fundación TikTok contaba con unos 55 millones de usuarios, pero su evolución ha sido extraordinaria y, tan sólo 2 años después, en 2020 llegó a superar los 689 millones de usuarios registrados en la red social. Uno de los factores a los que se atribuye este gran aumento es la gran cantidad de personas que se vieron obligadas a quedarse en casa encerrados por la pandemia mundial del COVID-19, aumentándose la demanda de servicios online para satisfacer la necesidad de entretenimiento (Camacho, Cardozo y Cristancho, 2022; Catalina, 2021).

Imagen 1.1: Características de cómo surge TikTok.

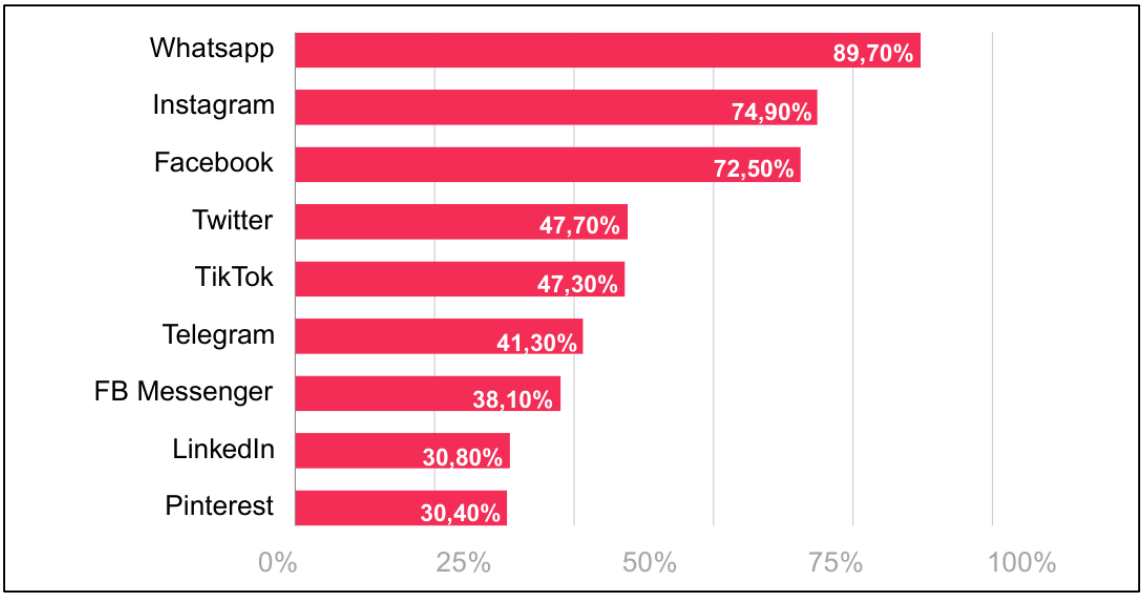


Fuente: Elaboración propia.

De hecho, TikTok fue la segunda red social con más descargas durante el primer trimestre del 2020 y se convirtió en ese mismo año, según Hootsuite, en la sexta red social más utilizada a nivel mundial. Además, ésta es la primera plataforma social con presencia global que no fue creada en EEUU, sino que su origen surgió en China (Larrondo, Morales y Peña, 2022; Herranz, Moya y Sidorenko, 2021).

Según el informe Digital 2023 Spain Report de We Are Social, TikTok se encuentra entre las cinco plataformas más utilizadas en España. Esto quiere decir que TikTok se sitúa como una de las redes sociales principales y, como se comentaba anteriormente, no pertenece a Meta como las tres más utilizadas.

Imagen 1.2: Redes sociales más usadas en España (enero 2023).



Fuente: Elaboración propia a partir de los datos del informe Digital 2023 Spain Report (We Are Social).

Hoy en día, TikTok ha conseguido afianzarse como una de las redes sociales más conocida a pesar de su poco tiempo activa si se compara con la antigüedad de las demás. Teniendo en cuenta el Estudio de Redes Sociales 2023 de IAB, cuando se da un conocimiento espontáneo sobre las redes sociales, TikTok es la cuarta red social que la gente tiene en cuenta. La plataforma llega a posicionarse por encima de YouTube, WhatsApp y LinkedIn. Cabe destacar que también ha sido la que más incremento a reflejado comparando con el estudio del año anterior.

Su gran impacto se debe principalmente a su novedosa forma de trabajo, centrándose principalmente en vídeos cortos verticales, filtros, *challenges* musicales y humor. La red social permite ver y compartir vídeos de una manera rápida y desenfadada gracias a su interfaz de contenido simple, llamativo e intuitivo. Aparte de esto, también permite categorizar vídeos, añadir etiquetas o *hashtags* para viralizar el contenido, posicionar a través de *me gustas*, etc., entre otras muchas funciones. Dentro de la red social se puede encontrar gran variedad de temáticas, pero la más utilizada es el humor, que está presente en gran cantidad de vídeos. Principalmente se puede resumir en dos pilares las razones del éxito de TikTok: lo sencillo que es su uso y lo viral que pueden llegar a ser sus vídeos (Larrondo, Morales y Peña, 2022; Martín y Micaletto, 2021).

El triunfo de esta red social ha cambiado el mundo de las redes sociales y por eso es tan importante descubrir cómo funciona la plataforma y qué la hace tan diferente a las demás.

1.2. AUGE DE LOS VÍDEOS CORTOS

TikTok destaca por fomentar la creatividad y el talento de los usuarios, ya que les permite mostrarse tal cual son a través de la creación de vídeos cortos y sencillos. Los usuarios de esta red social, denominados *TikTokers*, suben a sus perfiles los vídeos que ellos mismos editan y producen a través de la propia interfaz de TikTok, expresando sus opiniones y mostrando sus dotes de *performance*. Por esta razón, la propia autoproducción de los usuarios estimula las habilidades interpretativas e incluso la práctica de mezclar diferentes sonidos con efectos visuales. Asimismo, diferenciándose de otras redes sociales, no hay manera de hacer crecer el *feed* de un perfil si no se crea contenido, no existe forma de compartir publicaciones de otros perfiles (Conde, 2021, García y Suárez, 2021; Ballesteros, 2020). Esto lleva a que, si no se crea contenido propio, el perfil del usuario estará vacío.

El principal objetivo de todas las redes sociales es que los usuarios inviertan el mayor tiempo posible dentro de la plataforma, y para ello el contenido que se ofrezca tiene que llamar la atención de los usuarios. Este es el motivo por el cual TikTok promueve el dinamismo a través de la rapidez de creación de contenido e innovación creativa constante. No obstante, durante la pandemia mundial del COVID-19, la época de auge de la plataforma, realmente no se ha generado algo útil, sino que ésta ha mantenido a la gente entretenida con humor y vídeos intrascendentes (Herranz, Moya y Sidorenko, 2021; Méndez y Olivares, 2020).

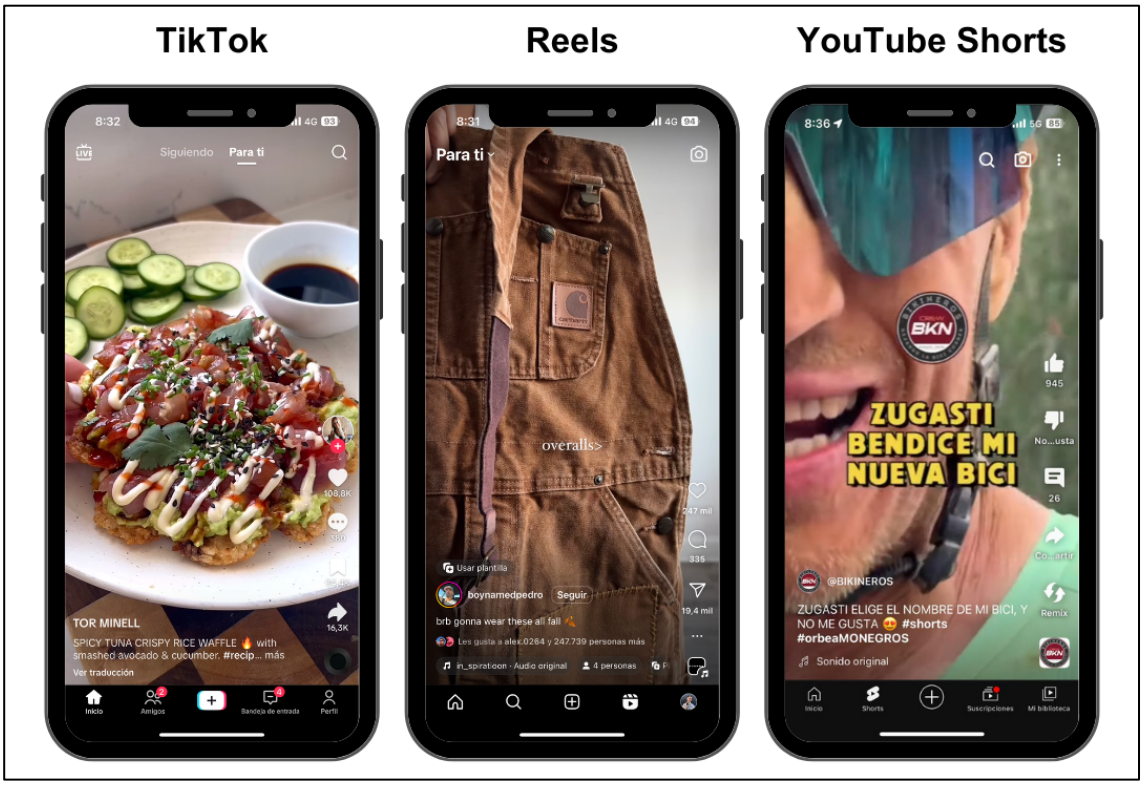
TikTok explota bastante bien este aspecto, fomentando la adicción a sus contenidos gracias a que el algoritmo que utiliza personaliza el contenido que se muestra a cada usuario en función de sus gustos y opiniones. Además, dentro de la plataforma se utiliza el formato más idóneo para enganchar a los usuarios: el formato de pantalla completa, que abstrae de cualquier otro estímulo que pueda causar el smartphone (Herranz, Moya y Sidorenko, 2021; Méndez y Olivares, 2020).

Desde que el smartphone se impuso como instrumento de comunicación y entretenimiento, los contenidos que se crean se han transformado en formatos verticales. De la misma forma, no se requiere mucha atención para la visualización de este tipo de vídeos gracias al propio formato del contenido audiovisual, mensajes breves y simples. Por otra parte, si el vídeo no gusta al usuario lo único que tiene que hacer es deslizar hacia arriba y ya se encuentra viendo otro vídeo diferente, haciendo que sea mucho más adictivo y se invierta más tiempo del pensado (García y Salvat, 2022; Ballesteros, 2020).

Con el triunfo de TikTok, los vídeos móviles se han convertido en la tendencia principal de las redes sociales y potencialmente en un formato breve para mantener la atención del espectador. La principal razón del auge de este tipo de formato va relacionada con lo comentado anteriormente, el consumo de redes sociales va en aumento y cada vez más se utiliza el smartphone cuando los usuarios se encuentran en movimiento. Esto quiere decir que el contenido debe de estar estructurado y producido para que sea rápido de visualizar, lo que se llama *fast content*. Los vídeos cortos son el tipo de contenido más adecuado para cubrir la necesidad de lectura rápida y fragmentada (Bahiyah, 2020; Ballesteros, 2020).

De hecho, a raíz del triunfo de TikTok y el auge de este tipo de vídeos, las demás redes sociales más conocidas han intentado integrar en su interfaz los vídeos cortos verticales. En el caso de Instagram y Facebook ha sido la incorporación del apartado *reels* o, por ejemplo, en YouTube se incluyeron en la plataforma los *YouTube Shorts*.

Imagen 1.3: Visión general de TikTok, Reels y YouTube Shorts.



Fuente: Elaboración propia.

En la imagen anterior se aprecia que los apartados de vídeos cortos que han incorporado las demás plataformas son prácticamente idénticas a la interfaz de TikTok. Como se comentaba anteriormente, los vídeos ocupan toda la pantalla para que no exista otra distracción y, además, se utilizan tonos en negro para que el menú sea más disimulado. Esto denota que TikTok ha impuesto los vídeos cortos y ha marcado una nueva tendencia en las redes sociales.

1.3. VIRALIDAD Y SU ALGORITMO

Parte del triunfo de TikTok es su innovador sistema recomendando vídeos. Gracias al algoritmo que utiliza la red social, no solo se muestra contenido a los usuarios a través de un *feed* convencional, se muestra el contenido en función del vídeo y no siempre del emisor en sí, como podría ser un influencer. Esto hace que cualquier usuario pueda llegar a viralizarse sin necesidad de tener gran cantidad de seguidores. Al mismo tiempo, este hecho produce un sentimiento gratificante a los usuarios que llegan a tener un alto alcance en sus vídeos y se sienten una celebridad por un momento, haciendo que se impulse más todavía la creación de contenido (Herranz, Moya y Sidorenko, 2021; Ballesteros, 2020).

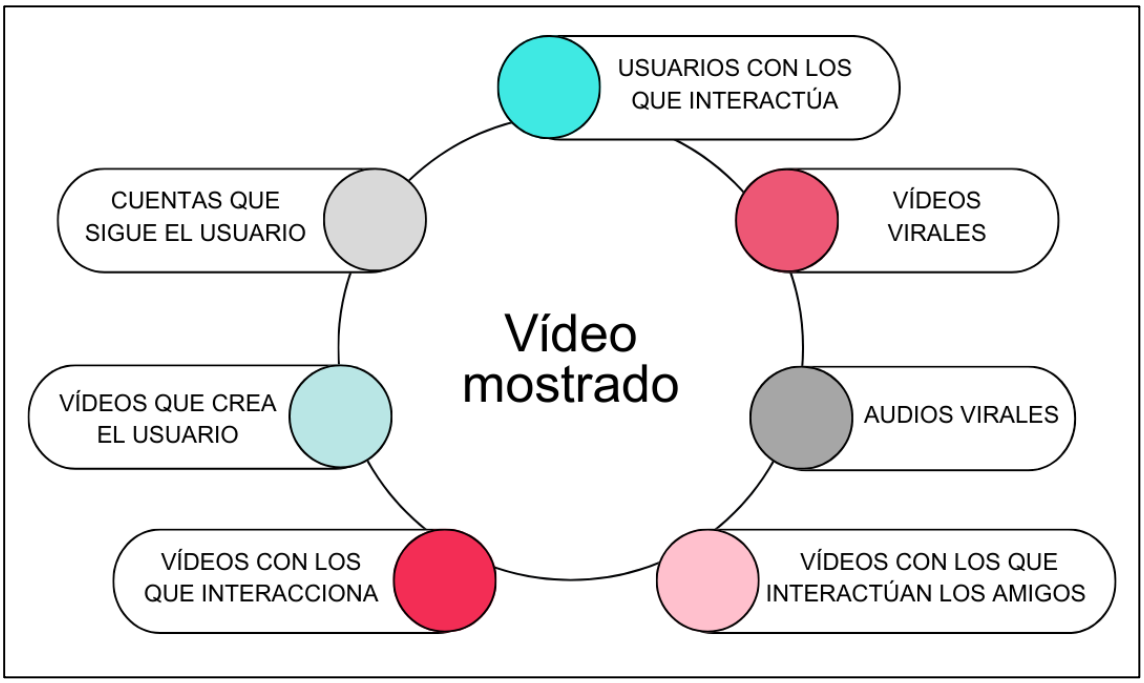
No obstante, no existen unos factores concretos que hagan que el contenido se convierta en viral, eso sí, en el caso de TikTok son el aspecto emocional y la capacidad

de proyección los que destacan entre todos los contenidos virales. Otro factor que puede facilitar o impulsar el aumento de visualizaciones es la utilización de *hashtags* y realizar los desafíos o *challenges* que estén en tendencia (Ballesteros, 2020; Méndez y Olivares, 2020).

De todos modos, con la aparición de los llamados *TikTokers* se favorece la viralidad de los vídeos generando cada vez más contenido e interacción entre los usuarios. Como es de esperar, los vídeos que se hacen virales promueven una mayor participación ya que obtienen más comentarios, me gustas y visualizaciones (Conde, 2021; Bahiyah, 2020).

Otra novedad que tiene TikTok frente a las demás redes sociales es que el usuario visualiza el contenido tomando decisiones instantáneas o intuitivas que hace en ese mismo momento, lo que rompe con el modelo tradicional de ver contenido basado en seguidores. La plataforma es capaz de aprender en función de los gustos y de los hábitos del usuario, y muestra un contenido diferente a cada persona. El algoritmo recoge información de los vídeos con los que el usuario interactúa, de los vídeos que va creando y de las cuentas que sigue para mostrar ese contenido adaptado (Larrondo, Morales y Peña, 2022; Martin, Merino y Micaletto, 2022).

Imagen 1.4: Aspectos que influyen a la decisión del vídeo mostrado en TikTok.



Fuente: Elaboración propia.

Por otra parte, también recaba información sobre qué vídeos están generando más *me gustas*, comentarios y visualizaciones en la plataforma. Esto se consigue gracias a la inteligencia artificial que incorpora, haciendo que la página de "Para ti" (página de inicio) llegue a provocar un efecto adictivo y que el usuario esté usando la aplicación por un tiempo más largo del que pretendía. No solo se busca que el usuario visualice contenido de los perfiles a los que sigue, no necesariamente tienes que ser seguidor de un perfil para ver sus vídeos. De hecho, la gran mayoría de los vídeos que el usuario encontrará en su página de inicio serán de usuarios a los que no está siguiendo. Esto hace que se cree el espacio inmersivo que se comenta, en el que la

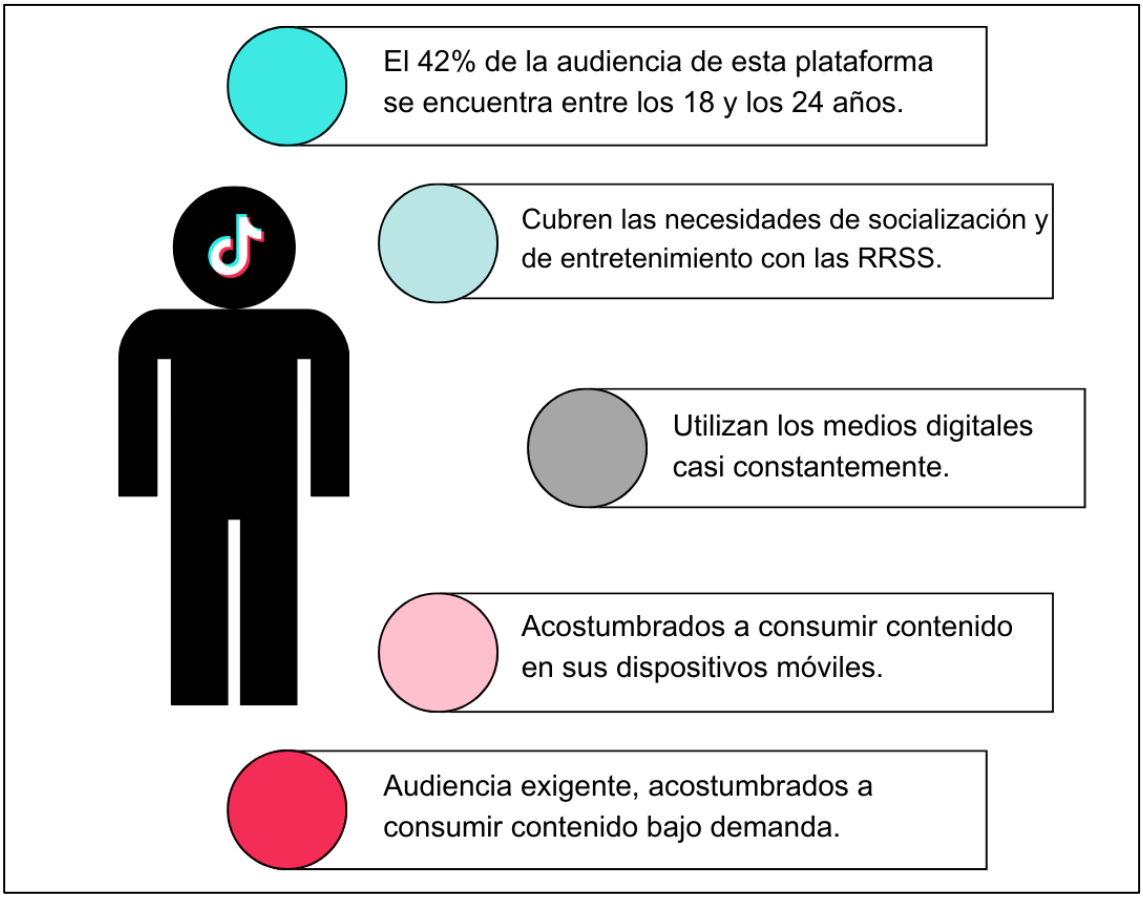
plataforma capta la atención del usuario por un tiempo prolongado (García y Salvar, 2022; Elhai, Montag y Yang, 2021).

Un dato a tener en cuenta a la hora de viralizarse es que no hace falta completar un registro para visualizar contenido de la plataforma, lo que hace llegar a los usuarios a una mayor audiencia. Este hecho hace que el usuario que sube el vídeo puede llegar a obtener visualizaciones incluso de gente que no está registrada en la red social, haciendo que su probabilidad de conseguir más número de visualizaciones aumente. No obstante, esto hace que existan más usuarios pasivos, es decir, que solo se dedican a mirar y navegar. Este tipo de consumo pasivo predomina en la plataforma, y hace que existan muchos usuarios que no producen contenido pero sí visualizan el de los demás, convirtiéndose en meros observadores (Martin, Merino y Micaletto, 2022; García y Suárez, 2021).

1.4. LA RED SOCIAL FAVORITA DE LA GENERACIÓN Z

Cada generación se define principalmente a través de los hábitos y los consumos que los caracteriza, y las generaciones más jóvenes han convertido a las redes sociales en un elemento principal de sus vidas. Además, cuando se habla de que las redes sociales son parte de la vida diaria de estas generaciones, es porque con ellas cubren las necesidades de socialización y de entretenimiento (Benavides, Feijoo y Pave, 2023).

Imagen 1.5: Principales características de la audiencia de TikTok.



Fuente: Elaboración propia.

Desde que se creó TikTok en 2018, la expansión de la plataforma no ha parado de crecer, llegando a convertirse en una de las plataformas favoritas de la Generación Z. La mayoría de los usuarios que utilizan TikTok se ubican mayoritariamente en la generación Z, a pesar de que en 2020, coincidiendo con la pandemia mundial del COVID-19, el número de usuarios millenials aumentara considerablemente (Larrondo, Morales y Peña, 2022; Martín y Micaletto, 2021).

La generación Z se puede considerar como la primera generación de usuarios completamente móviles y ha encontrado en TikTok lo mismo que en su momento los millenials hallaron en Facebook. El 42% de la audiencia de esta plataforma se encuentra entre los 18 y los 24 años, y a su vez, el 27% tiene una edad entre los 13 y los 17 años. Este dato refleja que la principal audiencia de esta red social representa perfectamente a la generación Z (García y Salvat, 2022).

Cabe destacar que el registro a la plataforma está permitido a partir de los 13 años, pero con la intención de proteger a los usuarios más jóvenes, la función de mensajes directos solo se activa a partir de los 16 años (Elhai, Montag y Yang, 2021). Hay que tener en cuenta que al registrarse se está concediendo a la empresa datos personales del usuario, como el teléfono móvil y/o correo electrónico, localización y contactos. En cuanto a la privacidad, todo el contenido que se sube a la red social se analiza y se almacena en su base de datos (Herranz, Moya y Sidorenko, 2021; Martín y Micaletto, 2021).

Los usuarios ubicados en esta generación están acostumbrados a consumir contenido en sus dispositivos móviles y muestran una mayor preferencia por el contenido de entretenimiento. TikTok permite a los jóvenes difundir contenido original proyectando una imagen determinada de ellos mismos (Martin, Merino y Micaletto, 2022).

La red social invita constantemente a la participación por parte de los usuarios con todas las tendencias que se van creando dentro de ella. Otro punto a favor de la plataforma es que, al ser usuarios principalmente jóvenes, son el colectivo que más utiliza los medios digitales llegando a utilizarlo constantemente y viendo las redes sociales como la manera de mantenerse en contacto con gente. Este hecho hace que la actividad de los usuarios en la plataforma sea mayor y, por tanto, TikTok disponga de más usuarios activos (Alonso, Giacomelli y Sidorenko, 2021; Conde, 2021).

Al mismo tiempo, esta generación está acostumbrada a consumir contenido bajo demanda, haciendo que sea una audiencia exigente. El contenido que consumen y la publicidad, les gusta que les genere un valor añadido y no que sea repetitiva, excesiva y que interrumpa su actividad en redes sociales (Benavides, Feijoo y Pavez, 2023). Básicamente, no quieren sentir que, mientras están activos en la red social, se les esté constantemente intentando vender productos o servicios por parte de las cuentas de las marcas.

1.5. PUBLICIDAD EN TIKTOK

El hecho de que las marcas se hayan ido introduciendo en las redes sociales se traduce como una búsqueda de crear una comunidad y de generar conversación con sus seguidores. Muchas de ellas, en sus perfiles, se encargan principalmente de resolver dudas y de ofrecer atención al cliente. En sus perfiles no solo pueden ofrecer y anunciar sus servicios y/o productos a los usuarios, sino que tienen que acercarse al consumidor y brindarle atención al cliente y manejar sus quejas y reclamos (Benavides, Feijoo y Pavez, 2023; Camacho, Cardozo y Cristancho, 2022).

Las redes sociales han marcado un cambio en la difusión de información de las empresas y ofrecen un nuevo diálogo entre empresa y consumidor. El usuario que recibe la información ya no es un mero lector, sino que también participa e interactúa con el contenido de la empresa, incluso llegando a generar más información de la marca (Maldonado, Oswaldo y Romero, 2023).

Por este motivo, no se puede negar que las redes sociales han despuntado convirtiéndose en un escaparate perfecto para realizar acciones promocionales, por ello, el triunfo de TikTok como una de las redes sociales más utilizadas también ha afectado a la manera de comunicar de las empresas. Las marcas que quieren conectar con su público tienen que adaptarse al universo de cada red social y cambiar el tono de su contenido publicitario (Martin, Merino y Micaletto, 2022).

TikTok es una plataforma complicada para las empresas debido a que tienen que adaptarse al tipo de contenido de la red social, lo que quiere decir, abandonar propuestas formales y publicidad tradicional. En definitiva, el contenido publicitario que realice una marca en su perfil de TikTok tiene que presentar un tono entretenido y no buscar exclusivamente vender, sino acompañar al usuario para estrechar una relación con él (Herranz, Moya y Sidorenko, 2021).

En las generaciones más jóvenes los mensajes publicitarios que se realizaban tradicionalmente y con un gran enfoque comercial, ya no llaman la atención ni generan el mismo efecto en los consumidores. Cuando un usuario se encuentra un perfil de una marca en una red social se establece otro tipo de comunicación con un toque más desenfadado y menos corporativo. Las empresas se encuentran con el desafío de no adaptar simplemente el mensaje publicitario, sino de impulsar la participación del usuario con la marca y de reinventar sus estrategias hacia enfoques más participativos y menos intrusivos. Además, la visión de los clientes hacia la marca es más positiva desde las redes sociales que si llegasen a la empresa por un medio tradicional. Estas nuevas acciones llevan a la fidelización de las personas e impactan positivamente en las decisiones de compra de los consumidores (Benavides, Feijoo y Pavez, 2023; Maldonado, Oswaldo y Romero, 2023).

Imagen 1.6: TikTok for Business.

Fuente: Elaboración propia.

Actualmente, los contenidos más persuasivos son los vídeos en primera persona contando la experiencia al resto de usuarios, y la red social más destacada para este tipo de contenido es TikTok (Conde, 2021). Los usuarios perciben que sus opiniones sobre productos o experiencias que vivieron con algunas empresas, tienen mayor cabida en esta red social e, incluso, mayor repercusión debido a la facilidad de viralización que se comenta anteriormente.

Las marcas cada vez invierten un mayor presupuesto en la publicidad en redes sociales frente a los medios tradicionales, es por eso que surgió la necesidad de crear una plataforma para las marcas. Como solución a esta nueva necesidad, TikTok desarrolló una plataforma de contratación de publicidad llamada *TikTok for Business*. La diferencia más notable frente a otras plataformas es que no es necesario tener un perfil activo para publicitarse, es decir, la marca no necesita tener un perfil en la red social para anunciarse dentro de ella (Camacho, Cardozo y Cristancho, 2022; Herranz, Moya y Sidorenko, 2021).

Dentro de esta plataforma, la marca puede crear los diferentes tipos de anuncios que la plataforma tiene disponibles, como por ejemplo, aparecer como primer vídeo cuando un usuario inicie el app, promocionar vídeos subidos en el perfil corporativo o crear anuncios que aparecerán en la sección de *Para Ti* de los usuarios. Otra novedad que tiene esta plataforma es que también ofrece el servicio de poner en contacto a la marca con creadores de contenido que encajen con la empresa. Además, en toda la plataforma se ofrecen consejos y recomendaciones en función de cómo sea la marca con el objetivo de facilitar el proceso.

1.6. EL MAL USO DE LA RED SOCIAL

La integración de las redes sociales en la vida de las personas puede llegar a ser un arma de doble filo en cuanto a la veracidad de lo que se puede leer en ellas. Cualquier persona con un perfil en redes sociales puede expresar su opinión, por eso es importante diferenciar opiniones de informaciones contrastadas. Debido a esto, las redes sociales se pueden convertir en un canal de información falsa si se hace un uso erróneo.

Las noticias falsas u opiniones basadas en informaciones falsas pueden llegar a tener un mayor alcance que las informaciones verificadas si se están difundiendo a través de redes sociales. Como se comentaba anteriormente, dentro de TikTok hay una mayor probabilidad de convertirse en viral y de alcanzar un mayor número de personas, lo que quiere decir que si una información falsa empieza a difundirse masivamente se está llegando a una desinformación de los usuarios (Alonso, Giacomelli y Sidorenko, 2021).

En este contexto en el que las redes sociales se han convertido en el medio para obtener información, se convierte en un reto diferenciar noticias de bulos. Si el foco se centra en TikTok, red social caracterizada por su comunidad joven, puede llegar a ser peligroso que se estén proporcionando constantemente informaciones que no son reales. Cuando un usuario está recibiendo estímulos frecuentemente sobre algún tema, y además es falso, se puede llegar a influenciar de manera negativa (García y Salvat, 2022).

Dentro de esta red social predomina el *fast content*, lo que hace que los primeros segundos del vídeo sean importantes para que el usuario decida instantáneamente si quiere verlo o no. Este hecho hace que la gente añada titulares llamativos en los

primeros instantes de los vídeos, llegando incluso a exagerar información que puede llevar a poca veracidad.

Destacar la importancia de la veracidad de la información que se difunde en TikTok es todavía más esencial cuando se trata de cuentas educativas. La gran diversidad de cuentas que existen en la plataforma hace que existan muchos usuarios dedicados principalmente a educar, como pueden ser temas de salud. Hay perfiles que su objetivo es concienciar a la gente sobre temas concretos o compartir información que puede ser relevante en el bienestar de las personas. Estos usuarios llegan a una audiencia principalmente joven, como se comentaba anteriormente, por eso es importante que no se extiendan bulos en asuntos que puedan llegar a afectar a las personas.

En definitiva, TikTok es un gran canal de difusión en el que se puede adaptar cualquier tipo de mensaje para llegar a un gran número de personas, pero es importante valorar la veracidad de las informaciones.

2. METODOLOGÍA Y OBJETIVOS

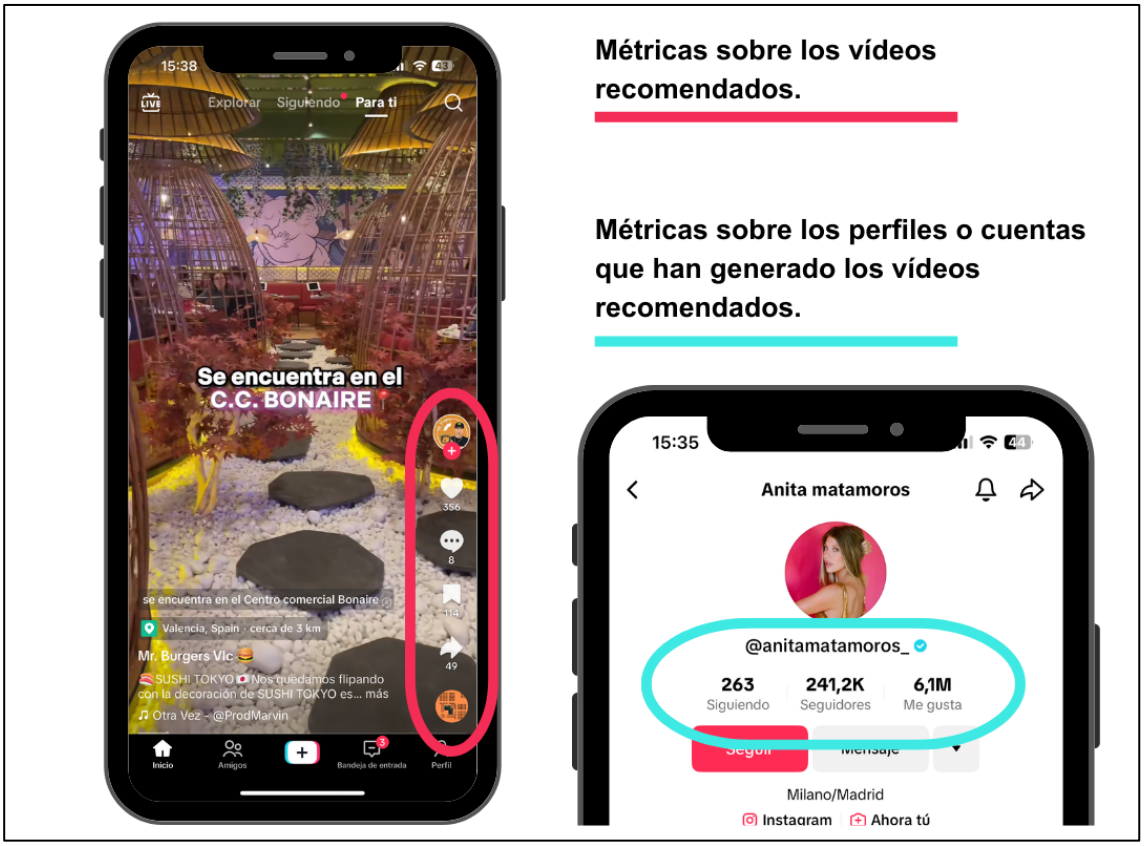
En relación con toda la información obtenida a través del marco teórico, surge el interés de realizar un estudio empírico que muestre cómo funcionan las recomendaciones de TikTok.

Como se expuso anteriormente, la gran diferenciación de la plataforma es la forma en la que se muestra el contenido hacia los usuarios y la utilización de inteligencia artificial en su algoritmo de recomendación. Por ello, se pretende analizar si el algoritmo de recomendación utiliza el sexo y la edad de los perfiles creados para realizar sus recomendaciones. Además, se valoran diferentes métricas, tanto de los vídeos recomendados como de los perfiles de las cuentas en los que éstos han sido generados, así como posibles relaciones entre ellas.

Con los datos recopilados en este estudio, se pretende determinar si el algoritmo de recomendaciones que utiliza TikTok en perfiles recién creados, y que no han realizado ningún *me gustas,* ni expresado preferencias, tiene algún tipo de sesgo respecto a la edad y el sexo del perfil creado, es decir, si el algoritmo tiene en cuenta estos atributos en las recomendaciones que realiza. Este es el objetivo principal de nuestro estudio.

Tomando como referencia la información previa, se ha llevado a cabo una recogida de datos creando doce perfiles ficticios en la plataforma de TikTok, controlando el sexo y la edad de los mismos. Posteriormente se recopilaron métricas sobre los diez primeros vídeos recomendados a cada uno de estos perfiles ficticios.

Imagen 2.1: Visión general de los vídeos y perfiles donde se obtienen las variables en TikTok.



Fuente: Elaboración propia.

Al controlar el sexo y la edad de los perfiles creados, nuestro estudio nos permitirá, además de estudiar las relaciones entre las diferentes métricas, determinar relaciones de causalidad entre los factores controlados y el resto de variables, ya que la forma en que se han generado los datos permite encuadrar nuestro estudio en el campo del diseño de experimentos.

Previamente al trabajo de campo se realizó la elección y definición de las variables a recoger, tanto de los vídeos recomendados como de los perfiles o cuentas de los mismos.

La recogida de datos se ha realizado en una hoja de cálculo Excel, (se adjunta en el *Anexo II*) lo que facilita el orden y la creación de una base de datos para su posterior análisis en el programa elegido.

Además del objetivo principal, descrito anteriormente, otros objetivos de esta investigación son:

- Estudiar las características que presentan tanto los vídeos recomendados por TikTok como las cuentas o perfiles que han generado estos vídeos.
- Estudiar la influencia de los factores controlados en el tipo de perfiles o cuentas al cual pertenecen los vídeos que el algoritmo recomienda para perfiles nuevos.
- Determinar si existe algún tipo de relación entre las características de los vídeos recomendados que TikTok muestra a través de su algoritmo.
- Determinar la influencia que la fecha concreta o la geolocalización de los dispositivos desde los que se han creado los perfiles tienen en el tipo de recomendaciones que reciben.
- Utilizar modelos de predicción para determinar características de los vídeos recomendados, en función del sexo y la edad del perfil al que se recomiendan dichos videos.

Hemos utilizado el programa estadístico R, mediante el entorno de desarrollo integrado RStudio, tanto para realizar el análisis exploratorio de nuestra base de datos, como para la obtención de los modelos de predicción, aplicando diferentes técnicas estadísticas según el tipo de variables y el objetivo perseguido.

Como hemos indicado anteriormente, el archivo de nuestra base de datos *tiktok.xlsx* recopila datos sobre algunas métricas de los diez primeros vídeos sugeridos para cada uno de los doce perfiles creados. Dentro de la base de datos se recoge información tanto de los vídeos sugeridos como de los perfiles o cuentas que han generado estos vídeos.

La base de datos consta de 120 filas y 25 columnas. Cada fila o caso de la base de datos recoge información asociada a cada vídeo sugerido por TikTok.

Las columnas de la base de datos recogen información sobre cada caso, y pueden clasificarse en tres grupos:

1. Factores controlados por los investigadores para crear los vídeos

    - *sexopc: sexo indicado en el perfil o cuenta ficticia que se ha creado*. Variable categórica con 2 niveles: *hombre, mujer*.

- *rangopc*: *rango de edad del perfil ficticio creado.* Recoge el rango de edad en el que se encuentra el perfil que se ha creado de forma artificial para conseguir los objetivos del TFM. Se han considerado los siguientes rangos de edad: de *15 a 18* años, de *19 a 25* años y de *26 a 60* años.

- *edadpc: edad del perfil ficticio que se ha creado*. Esta variable muestra la edad del perfil creado, se trata de una variable cuantitativa con los siguientes niveles de edad: 16, 18, 20, 24, 27 y 37 años de edad.

- *fnacpc: fecha de nacimiento del perfil ficticio creado*. La variable hace referencia a la fecha de nacimiento que se ha determinado a la hora de crear cada perfil. Puede tomar cualquier valor inferior o igual al año 2007. Se ha recopilado información sobre esta variable porque se solicita al crear los perfiles, pero no ha sido utilizada en los análisis realizados.

- *fcreacionpc: fecha en la que se ha creado el perfil ficticio*. Recoge la fecha en la que se ha creado el perfil y, por tanto, la fecha en la que se analizan los vídeos recomendados. Toma valores entre: 20-05-2023 y el 27-05-2023.

Considerar dos sexos, y seis valores diferentes para la edad de cada uno, da lugar a un total de 2x6=12 perfiles ficticios diferentes. Para cada perfil ficticio se han estudiado las 10 primeras recomendaciones, obteniendo así, un total de 120 casos (filas en nuestra base de datos).

2. Métricas sobre los perfiles o cuentas que han generado los vídeos recomendados

    - *perfil: identificador de la cuenta a la que pertenece el vídeo recomendado.* Número o cadena de caracteres que identifica al perfil del vídeo que ha aparecido como recomendado.

    - *sexo: sexo del perfil.* Determina el sexo del perfil o cuenta que ha enviado la recomendación, es una variable cualitativa nominal con cuatro categorías o niveles (*hombre, mujer, no binario, ninguno*). La categoría *ninguno* se asocia a perfiles o cuentas en las que no se distingue ni se identifica ningún sexo o pertenece a personas de distintos sexos. Asimismo, las demás variables también hacen referencia a perfiles que pertenecen a un grupo de personas de un mismo género, y no necesariamente solo a una persona.

    - *rango: rango de edad del perfil.* Recoge el rango de edad en el que se encuentra la persona del perfil que aparece en el vídeo recomendado. Se han considerado los siguientes rangos de edad: -18 = menos de 18 años, [18, 25] = entre 18 y 25 años, +25 = más de 25 años de edad. En los casos en que el perfil no muestre explícitamente la edad, se le asignará una edad orientativa basada en nuestra observación.

    - checkazul: tipo de verificación del perfil. Esta variable hace referencia a si el perfil del vídeo recomendado tiene el check azul o no. El check azul marca si el perfil de una persona con cierto reconocimiento es su

perfil oficial. Se trata de una variable cualitativa nominal con dos categorías (sí, no).

- *siguiendo: personas que sigue el perfil.* Número de cuentas que está siguiendo el perfil del vídeo recomendado en la parte de *para ti*. Se trata de una variable cuantitativa discreta ya que puede tomar cualquier valor entero mayor o igual a 0.

- *seguidores: número de personas que siguen al perfil.* Número total de seguidores que tiene la cuenta que envía la recomendación. Se trata de una variable cuantitativa discreta ya que puede tomar cualquier valor entero superior o igual a 0, como ocurre con la variable "siguiendo".

- *nmegustas: número total de "me gustas".* Esta variable mide el número total de me gustas que tiene el perfil. Es la suma de me gustas de todos los vídeos que ha subido este usuario desde que se abrió el perfil. Es una variable cuantitativa discreta ya que puede tomar cualquier valor entero superior o igual a 0.

3. Métricas sobre los vídeos recomendados

- nrecomendacion: orden en que aparece como vídeo recomendado. Esta variable refleja, dentro de los 10 vídeos recopilados, en qué orden ha aparecido. Variable cuantitativa con las categorías [1, 10].

- *sexvid: sexo del principal actor del vídeo.* Normalmente coincidirá con el sexo del perfil de la cuenta que lo envía. Variable categórica con tres categorías *(hombre, mujer, ninguno).* La categoría *ninguno* incluye vídeos donde no aparece ningún actor principal, o donde son varios y de distinto sexo, sin prevalecer ninguno.

- *mgvid: número de me gustas del vídeo.* Contabiliza el número total de *me gustas* que ha recibido el vídeo recomendado. Se trata de una variable cuantitativa discreta porque puede tomar cualquier valor entero mayor o igual a 0.

- *comentarios: número de comentarios del vídeo.* Determina el número de comentarios que ha obtenido el vídeo que aparece como recomendación. En este caso se trata de una variable cuantitativa discreta ya que toma cualquier valor entero mayor o igual a 0.

- *guardados: número de veces que ha sido guardado el vídeo.* La variable recoge cuántas veces ha sido guardado el vídeo por los usuarios que lo han visualizado, por tanto, es una variable cuantitativa discreta que puede tomar cualquier valor entero mayor o igual a 0.

- *compartidos: número de veces que ha sido compartido el vídeo.* Esta variable muestra cuántas veces ha sido compartido por parte de los usuarios que han visualizado el vídeo. Dicha variable es cuantitativa discreta y puede tomar cualquier valor entero mayor o igual a 0.

- *duracion: duración (en minutos) del vídeo.* La variable hace referencia a la duración en minutos del vídeo que ha aparecido en las recomendaciones, se trata de una variable cuantitativa continua. Puede contener cualquier duración entre 0 y 3 minutos, ya que ésta es la duración máxima de un vídeo que permite la plataforma. Finalmente no se ha recogido porque proporciona poca información y es muy complejo recabarla.

- *hot: contenido hot (+18).* Esta variable mide si el vídeo recomendado tiene contenido adecuado sólo para adultos. Se trata de una variable cualitativa nominal con dos categorías (*sí, no*).

- *tematica: temática general del vídeo recomendado*. Incluye la temática del vídeo que aparece como recomendado, es una variable cualitativa nominal. Las categorías consideradas han sido: *humor, baile, maquillaje, noticia en tendencia, provocativo, cocina, storytime, producto/unboxing* u *otro*. La categoría *storytime* hace referencia a cuando el vídeo trata de una persona contando una historia sobre algo que le ha pasado.

- tiporec: *tipo de recomendación.* Esta variable mide si el vídeo recomendado es o no un anuncio, es decir, si se trata de un vídeo de un perfil o si se trata de un vídeo de cualquier tipo de publicidad pagada a través de la plataforma. Teniendo en cuenta esto se trata de una variable cualitativa nominal con dos categorías (*perfil, anuncio*).

- *tipoaudio: tipo de audio.* Recoge si el vídeo recomendado tiene un audio original o incluye un audio reutilizado, es decir, un audio superpuesto al original. En este caso se trata de una variable cualitativa nominal que solo puede tomar dos valores (*original, superpuesto*).

- *nvidaudio: número de vídeos que utilizan el audio del vídeo recomendado.* Recoge el número de veces que se ha utilizado el audio usado en el vídeo en toda la plataforma de Tiktok. Esta variable es cuantitativa discreta y puede tomar cualquier valor entero superior o igual a 0.

3. RESULTADOS DEL ANÁLISIS EXPLORATORIO

3.1. Análisis descriptivo de las variables.

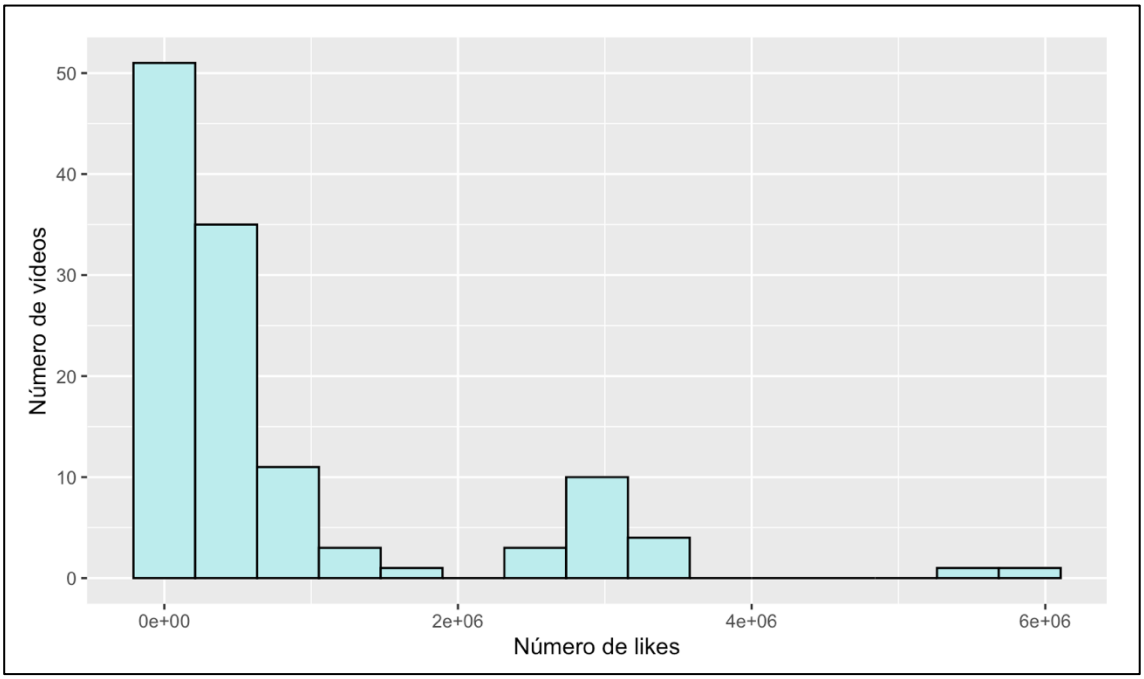
3.1.1. Métricas de los vídeos recomendados.

Me gustas de los vídeos recomendados.

Un primer análisis descriptivo de esta variable muestra la enorme variación que presentan los *me gustas* en los vídeos recomendados. Se puede observar que existe tres pequeños grupos de vídeos con un número de *me gustas* similares. Lo más destacable es que la gran mayoría de los vídeos están concentrado en el primero, que representa la parte de los vídeos con menos *me gustas*. No obstante, esto no quiere decir que se trate de vídeos con una cifra pequeña ya que el mínimo de la variable analizada es 5.868 me gustas. Los demás grupos contienen muchos menos vídeos, como es de esperar, porque se tratan cifras de *me gustas* muy elevadas.

La mediana es de 306.250, por lo que el 50% de los vídeos recomendados tienene más de 300 mil *me gustas*.

Figura 3.1: Histograma del número de me gustas de los vídeos recomendados.

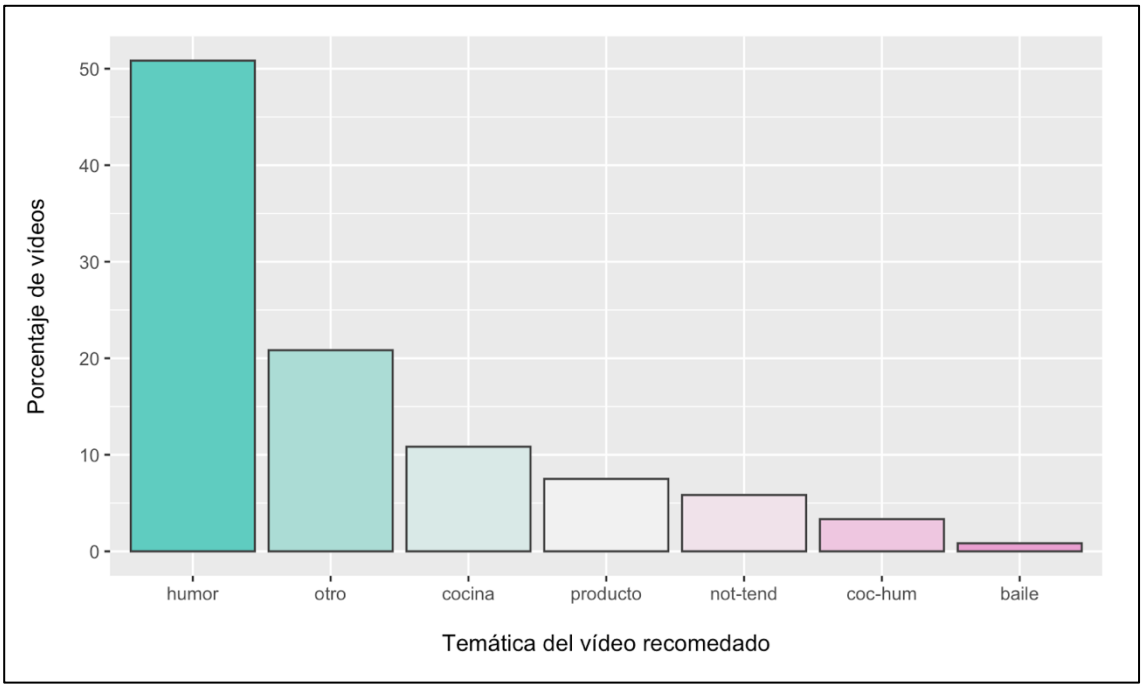


Fuente: Elaboración propia.

Temáticas de los vídeos recomendados.

Otra variable analizada ha sido la temática del contenido que se iba recomendando dentro de la plataforma. Observando las frecuencias de esta variable categórica, se puede ver que la temática más repetida ha sido el humor. Este dato sugiere lo que se ha comentado en el marco teórico: gran parte del triunfo de TikTok ha sido el humor de su contenido.

Figura 3.2: Gráfico de barras de la temática en frecuencias relativas.



Fuente: Elaboración propia.

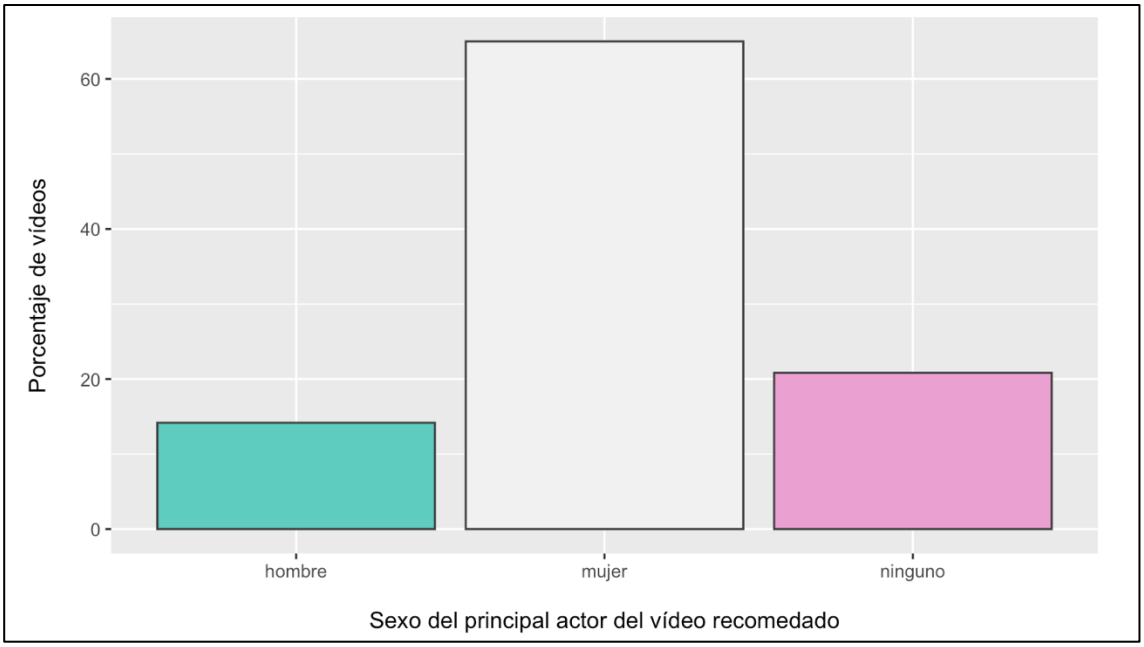
Si se tiene en cuenta las frecuencias relativas, el humor es la temática recogida en más del 50% de los vídeos recomendados. Es decir, la mayoría de los vídeos que TikTok ha sugerido a los perfiles ficticios tienen un tono humorístico. Un dato que destaca frente al marco teórico es la poca frecuencia de vídeos que tratan sobre bailes, ya que éste es uno de los pilares más importantes en esta red social. Además de la categoría que engloba a otro tipo de vídeos, destaca la categoría de vídeos relacionados con la cocina.

Sexo del actor principal que aparece en el vídeo recomendado.

Una de las variables a destacar es el sexo del actor principal que aparece en el vídeo recomendado. Como se ha comentado anteriormente, se han sugerido principalmente vídeos en los que el actor principal eran mujeres, por lo que existe una mayor proporción de vídeos de este tipo.

En la *Figura 3.3* se puede observar cómo las mujeres protagonistas de vídeos representan más del 60% de los vídeos sugeridos que se han recogido en la base de datos.

Figura 3.3: Gráfico de barras de las frecuencias relativas del sexo del actor principal del vídeo.

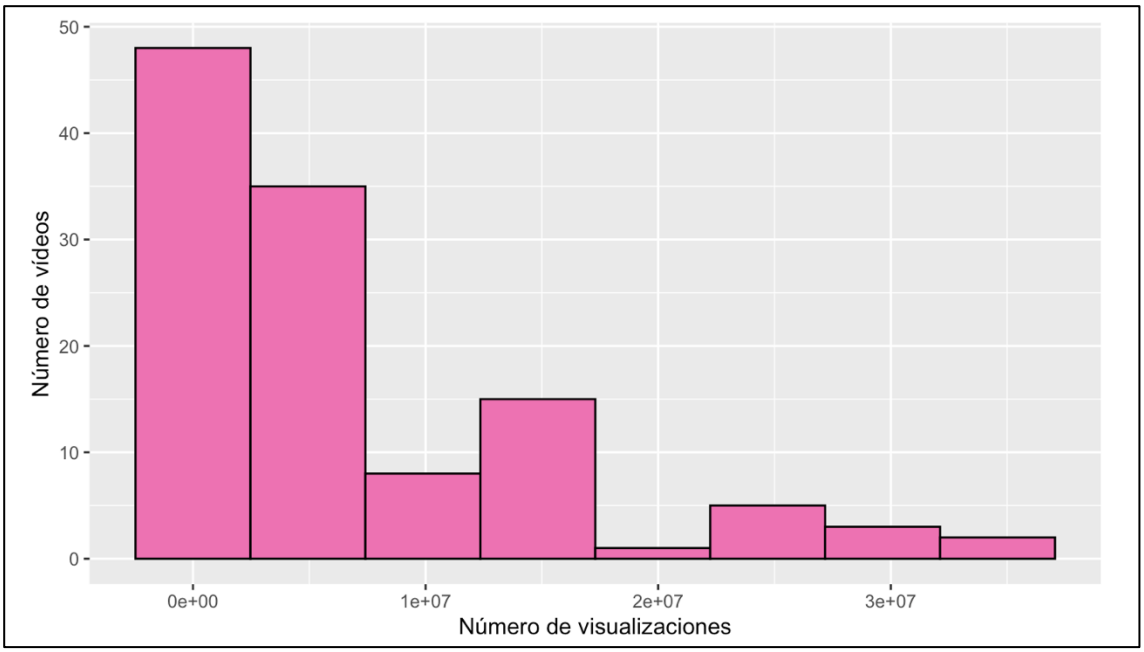


Fuente: Elaboración propia.

Visualizaciones.

En cuanto a las visualizaciones de los vídeos, ocurre al parecido a los me gustas. Los datos recogidos tienen mucha dispersión entre sí. No obstante, en este caso, se observa que la mayoría de los vídeos están en la zona de menores visualizaciones. Hay que tener en cuenta que el mínimo de esta variable se sitúa en 208.400 visualizaciones, lo que quiere decir que los vídeos mostrados son de grandes cifras de *views*.

Figura 3.4: Histograma de las visualizaciones de los vídeos recomendados.



Fuente: Elaboración propia.

Esta variable presenta una mediana de 3.200.000 visualizaciones. Este dato refleja la magnitud de los datos manejados, el 50% de los vídeos tienen más de 3.2 millones de *views*.

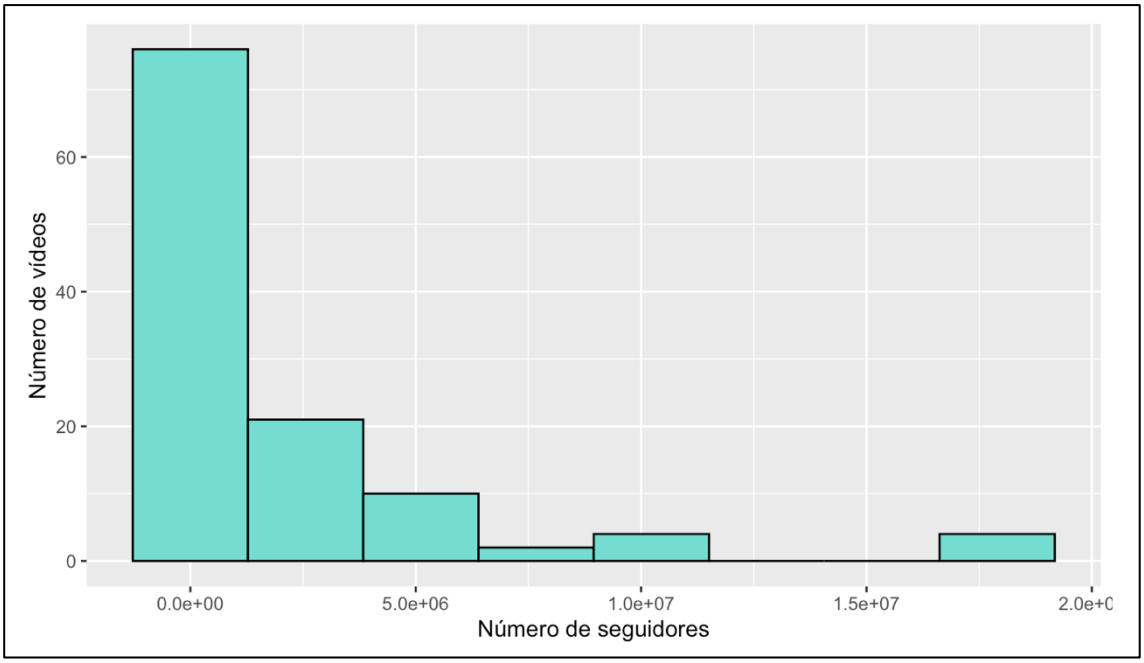
3.1.2. Características y métricas de las cuentas que han generado los vídeos recomendados.

Seguidores.

Como es de esperar, esta variable también presenta una gran dispersión entre los datos recogidos. En el marco teórico se comentaba que no era necesario tener seguidores para aparecer en el *feed* de las personas, así que el dato de que la mayoría de los perfiles se sitúen en las menores cifras de seguidores concuerda con lo comentado. Además, el mínimo de esta variable se sitúa en 670 seguidores, así que se verifica aún más esta teoría. El número medio de seguidores supera los 2 millones, aunque el 50% de los perfiles recomendados tienen menos de 519.400 seguidores.

Por otra parte, en la *Figura 3.5* también se observa que hay un pequeño número de perfiles que sí tienen grandes cifras de seguidores, con un máximo en torno a los 18 millones.

Figura 3.5: Histograma del número de seguidores de los perfiles de los vídeos recomendados.

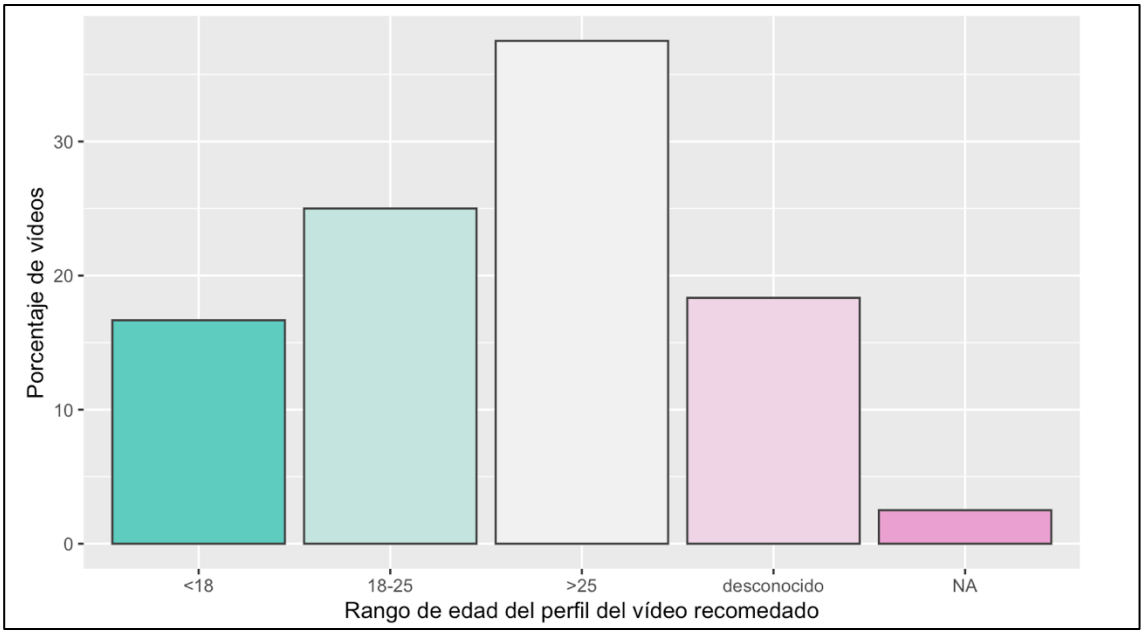


Fuente: Elaboración propia.

Rango de edad.

En la *Figura 3.6* se puede observar cuál es el rango de edad de los perfiles que se han mostrado en recomendaciones. Los perfiles que se estimaron como mayores de 25 años representan más del 35%. Un dato que destaca es la gran proporción de perfiles en los que la edad es desconocida con casi un 20%, esto se debe a que son perfiles de marcas o corporativos.

Figura 3.6: Gráfico de barras de las frecuencias relativas del rango de edad de los perfiles.

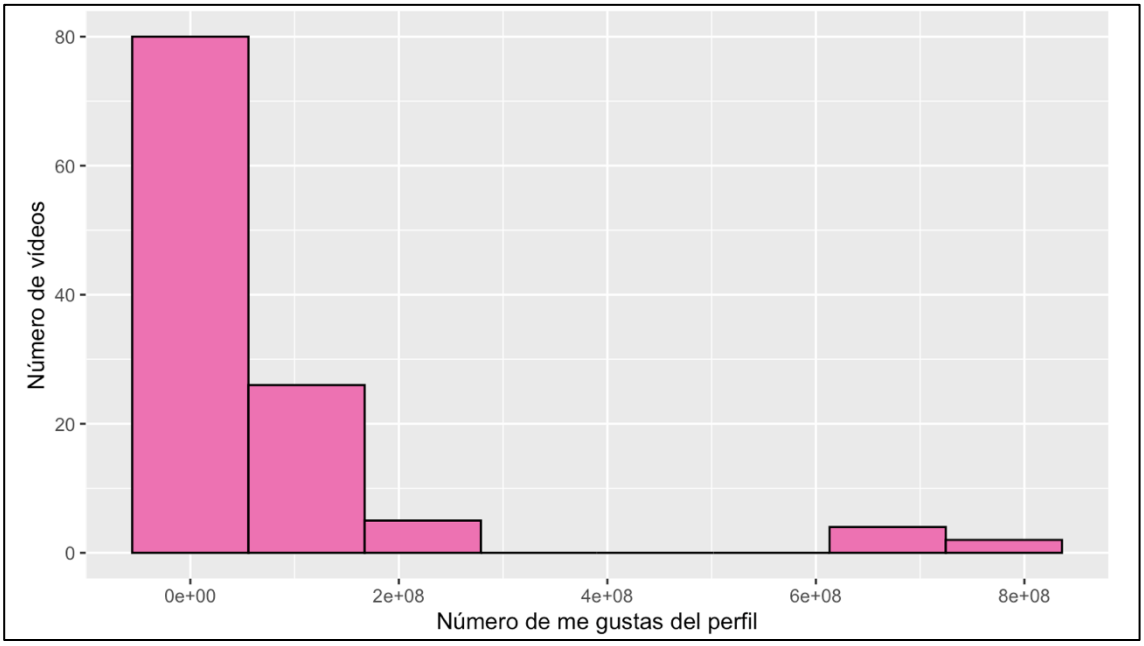


Fuente: Elaboración propia.

Número de me gustas del perfil.

Otro factor a tener en cuenta es el número de me gustas que tiene el perfil. En la Figura 3.7 se diferencian claramente los dos grupos de perfiles que existen en los datos. Hay un pequeño grupo de perfiles que tienen un número de me gustas mucho mayores a los del resto, pudiendo deberse a perfiles de personas con gran reconocimiento mediático.

Figura 3.7: Histograma del número de me gustas de los perfiles.



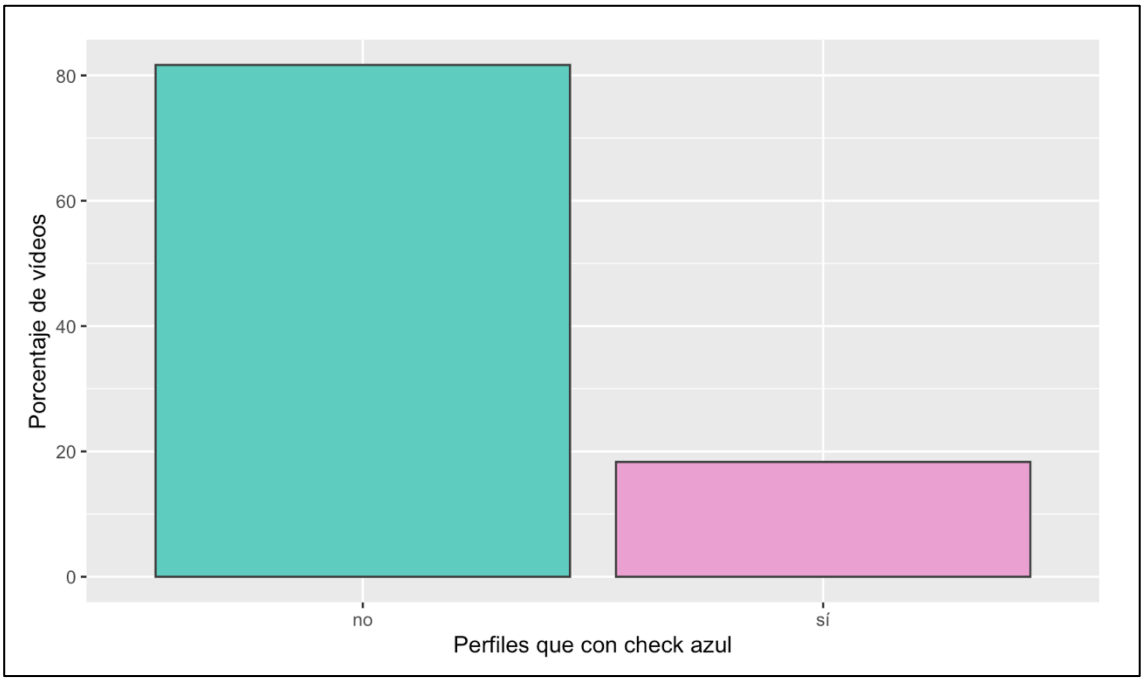
Fuente: Elaboración propia.

Check azul.

La verificación de los perfiles de TikTok puede mostrar que los perfiles tienen gran relevancia dentro de la red social. Dentro de los vídeos recomendados se puede observar que más del 80% de los perfiles mostrados no contaban con el *check azul*.

Este dato sugiere que, como se comentaba en el marco teórico, no se muestran los vídeos en función de la relevancia de su perfil. La gran mayoría de los vídeos recomendados no pertenecían a personas con gran influencia o relevancia dentro de la red social como para tener el verificado. En resumen, no tienes que ser un gran *influencer* reconocido para ser parte de los vídeos recomendados a perfiles nuevos.

Figura 3.8: Gráfico de barras de las frecuencias relativas del check azul.



Fuente: Elaboración propia.

3.2. Estudio de relaciones entre variables.

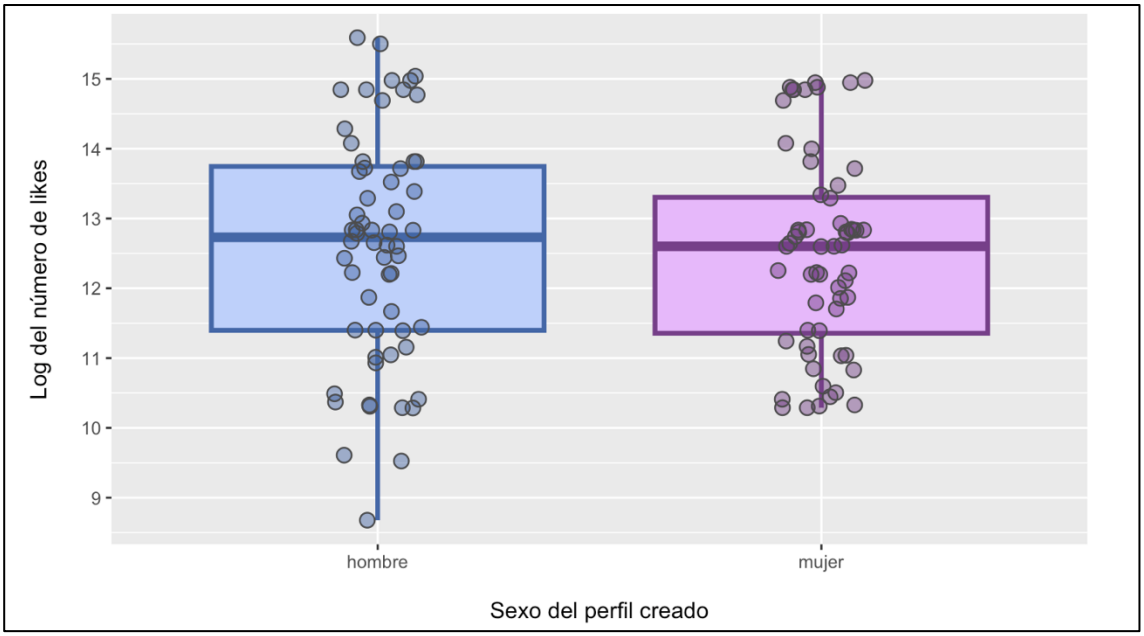
3.2.1. Entre métricas de los perfiles creados y otras métricas.

Sexo del perfil creado y los me gustas del vídeo recomendado.

El diagrama de cajas y puntos analizado no sugiere que existan grandes diferencias entre hombres y mujeres. Lo más destacable en cuanto a los datos recogidos en este gráfico es que la variación en el número de *likes* es mayor en los vídeos recomendados a hombres. Teniendo en cuenta los principales estadísticos de la distribución del número de me gustas para cada sexo, el rango de valores para las mujeres es inferior, y no tan amplio como en el caso de los hombres.

A través del análisis de la varianza (ANOVA) se obtiene que la diferencia no es significativa con un valor P=0.722.

Figura 3.9: Diagrama de caja del logaritmo del nº de me gustas según el sexo del perfil creado.



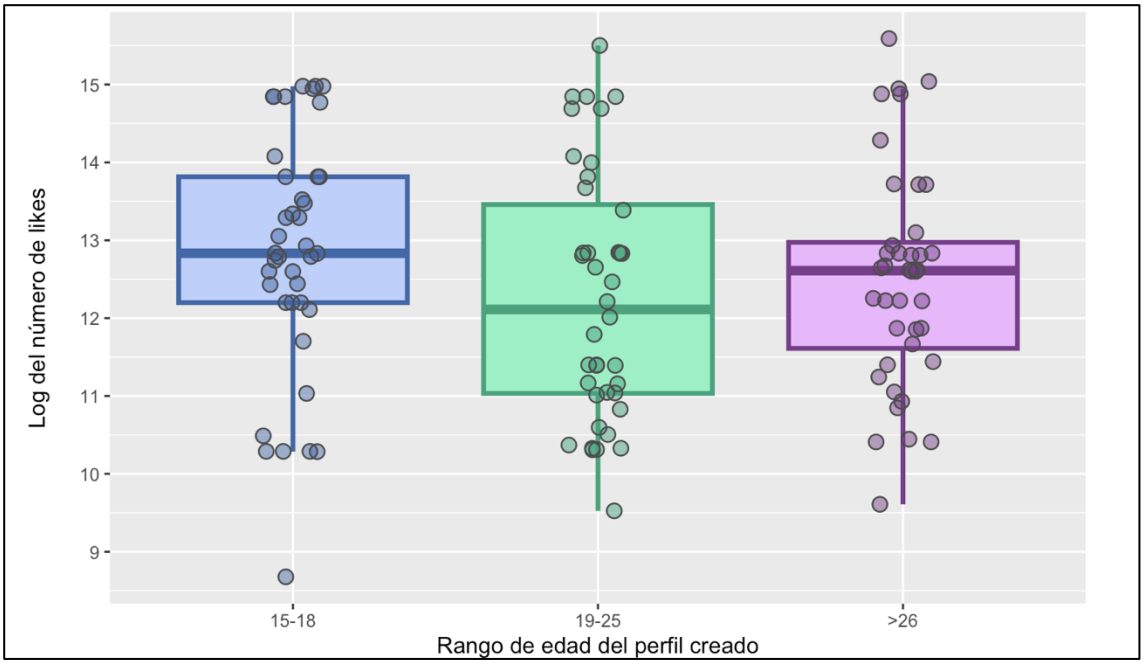
Fuente: Elaboración propia.

Rango de edad del perfil creado y los me gustas del vídeo recomendado.

Cuando se relaciona el número de *me gustas* y el rango de edad de los perfiles creados se puede observar en la *Figura 3.10* pequeñas diferencias pero ningún dato muy destacable.

Al realizarse el análisis de la varianza (ANOVA) se obtiene un valor P=0.25, lo que indica que no existen diferencias significativas entre los tres grupos analizados.

Figura 3.10: Diagrama de cajas del nº de me gustas (log) según el rango de edad del perfil creado.



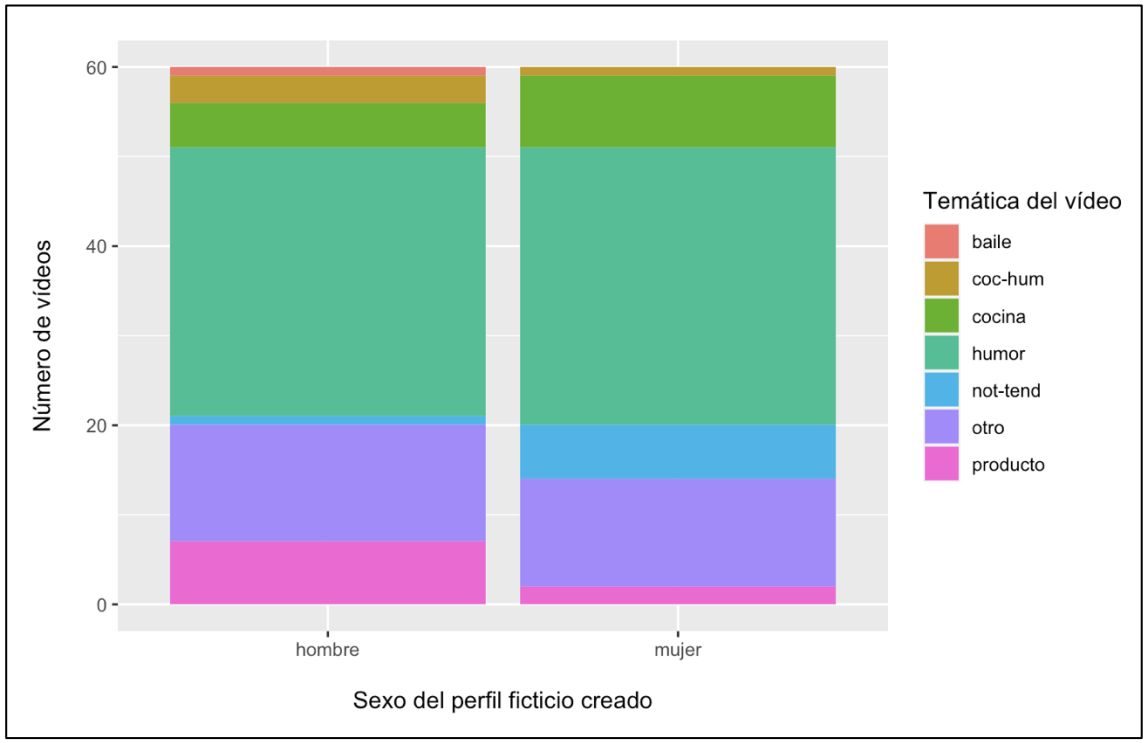
Fuente: Elaboración propia.

Sexo del perfil creado y la temática del vídeo recomendado.

Un aspecto importante en el análisis de esta variable es observar si TikTok sugiere diferentes temáticas de vídeos en función de si el perfil creado es hombre o mujer.

El gráfico de barras apiladas de la *Figura 3.11*. muestra pequeñas diferencias entre sexos. Las diferencias más notables para los perfiles de hombres es que se han sugerido menos vídeos de cocina y menos noticias en tendencia. Por otro lado, en los perfiles creados de mujeres aparecen en menor proporción los vídeos enfocados a mostrar productos. Un dato destacable es que los vídeos de baile solo aparecen cuando el perfil al que se muestran es un hombre.

Figura 3.11: Gráfico de barras de la temática según el sexo del perfil creado.



Fuente: Elaboración propia.

Para estudiar si las diferencias observadas eran significativas, utilizando un test de independencia Chi-cuadrado, tuvimos que agrupar algunas temáticas para conseguir que las frecuencias esperadas fuesen superiores a 5. La tabla con frecuencias cruzadas no muestra diferencias significativas (P=0.319, obtenido a partir de una prueba de Chi cuadrado para tablas de contingencia).

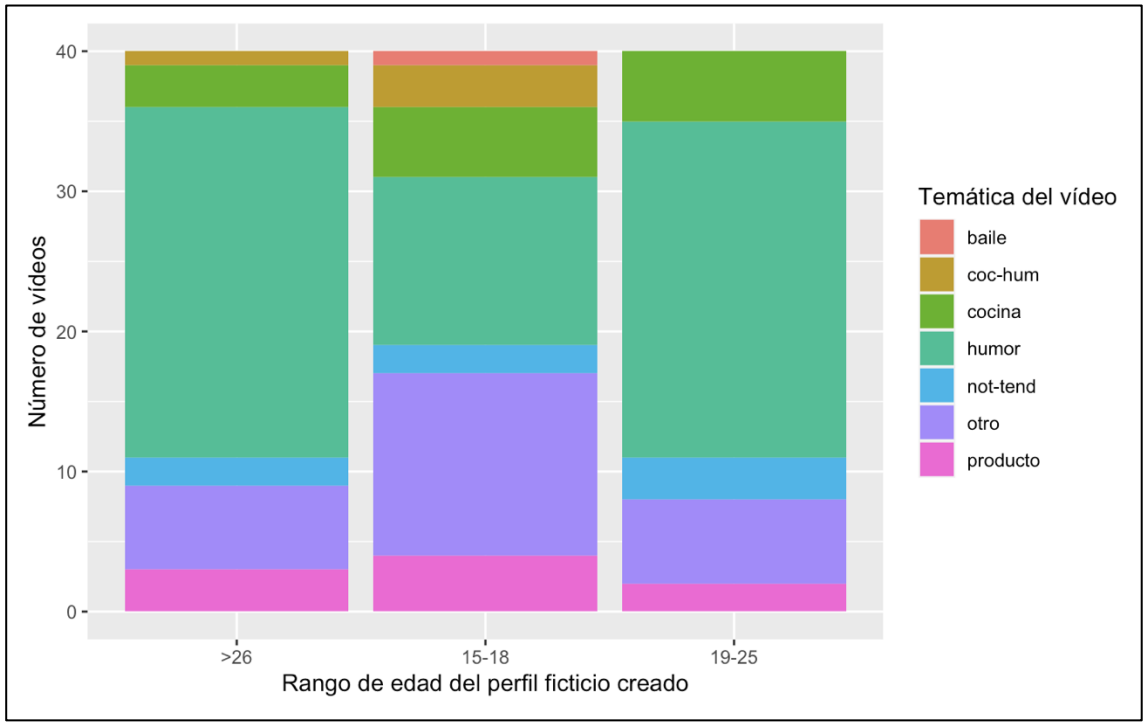
El estudio no permite concluir que el sexo del perfil creado sea un factor que determina la temática que presentan los vídeos recomendados.

Rango de edad del perfil creado y la temática del vídeo recomendado.

Por otra parte, cuando lo que se analiza es la relación entre la temática de los vídeos y el rango de edad del perfil creado, se pueden observar diferencias más notables a simple vista. Sobre todo, la mayor diferencia observable se encuentra en el rango de edad de 15 a 18 años. La temática en la que más cambio se aprecia es el humor, ya que en el rango de edad mencionado se reduce bastante su presencia al

compararlo con las demás edades. Además, en el caso comentado, aumenta la proporción de vídeos categorizados con la temática “otro” y “cocina y humor”.

Figura 3.12: Gráfico de barras de la temática según el rango de edad del perfil creado.



Fuente: Elaboración propia.

Como ha ocurrido en el anterior análisis de las variables, para realizar la prueba de significación a través de un test de independencia Chi-cuadrado se ha tenido que agrupar nuevamente categorías de la temática. En este caso, se han agrupado las mismas categorías para una mayor homogeneidad de los resultados.

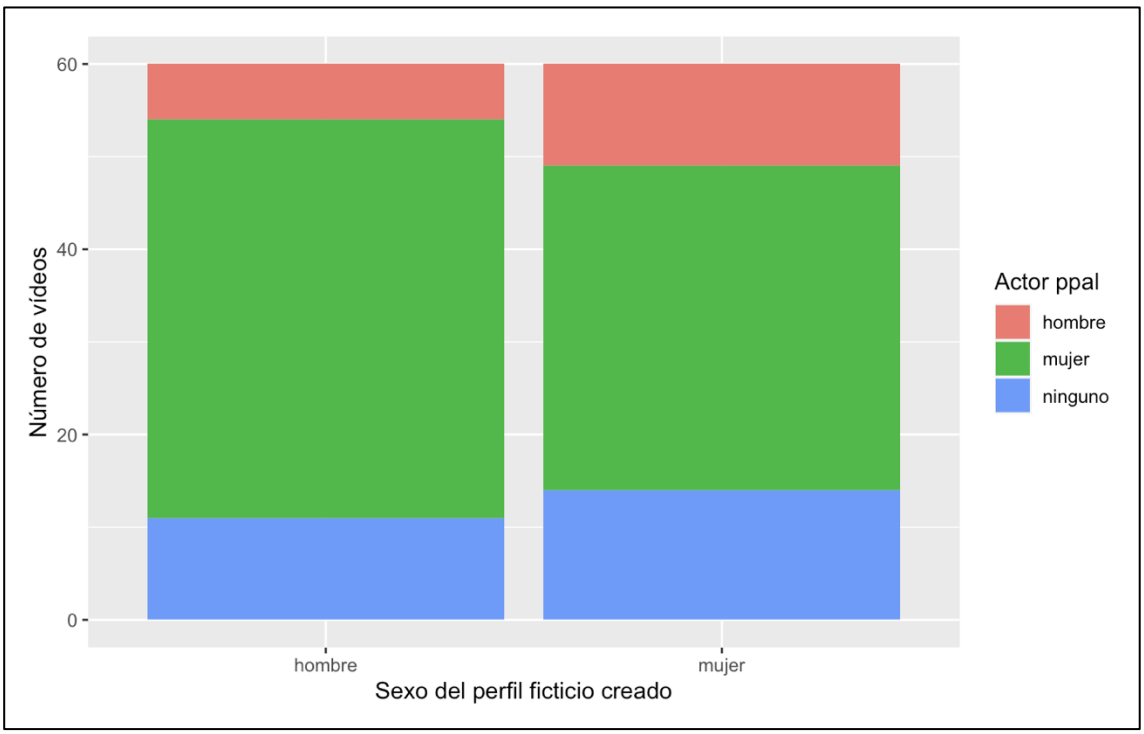
Con las categorías agrupadas se ha obtenido un valor P=0.065, por lo que las diferencias observadas son prácticamente significativas para un nivel del 5%.

Sexo del perfil creado y sexo del actor principal del vídeo.

Un análisis importante es comprobar si el sexo del actor principal del vídeo varía en función de si el perfil que lo está visualizando es el de una mujer o el de un hombre. En el gráfico de barras que se muestra a continuación, se aprecian pequeñas diferencias: cuando el perfil que está visualizando el vídeo es el de un hombre, la probabilidad de que el protagonista del vídeo recomendado sea una mujer aumenta.

En cambio, cuando la que visualiza el vídeo es una mujer se muestran más vídeos protagonizados por hombres o vídeos en los que no se identifica un sexo como protagonista del vídeo. Este último caso ocurre cuando en el vídeo no aparece ni una persona, ni grupos de personas en las que prevalezca un sexo.

Figura 3.13: Gráfico de barras de las frecuencias del sexo del actor principal del vídeo según el sexo del perfil creado.



Fuente: Elaboración propia.

Sin embargo, cuando se realiza la prueba Chi-cuadrado, para determinar si las diferencias observadas son significativas, obtenemos un valor P=0.27. El estudio no nos permite concluir, por tanto, que el sexo del vídeo recomendado esté influenciado por el sexo del perfil al que se dirige.

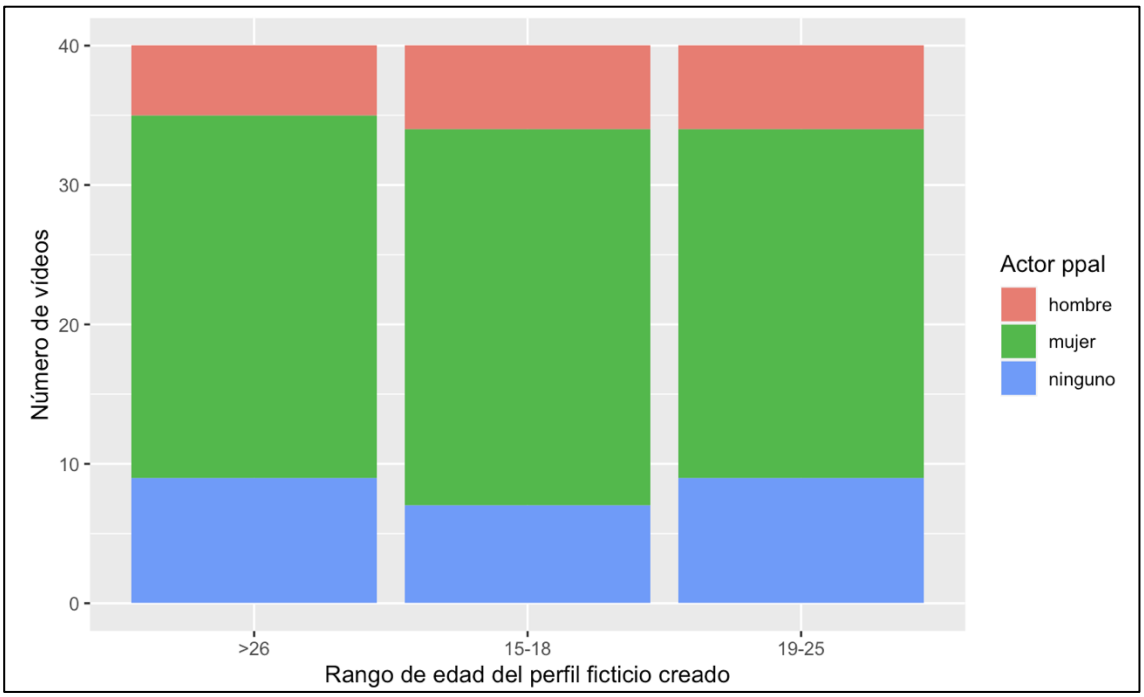
Rango de edad del perfil creado y el sexo del actor principal del vídeo.

Otras dos variables para las que es interesante analizar su relación son el rango de la edad del perfil creado y el sexo del actor principal de los vídeos. Si se observa la *Figura 3.14* a simple vista se puede ver que prácticamente existe la misma proporción para los tres rangos de edad definidos.

La prueba Chi-cuadro para comprobar si las variables son independientes muestra un valor P=0.97. Como era de esperar las diferencias que se obtienen no son significativas.

Este análisis lleva a la conclusión de que el rango edad de los perfiles creados no influye en si se muestra un vídeo protagonizado por una mujer, un hombre o ningún sexo identificado.

Figura 3.14: Gráfico de barras de las frecuencias del sexo del actor principal del vídeo según el rango de edad del perfil creado.

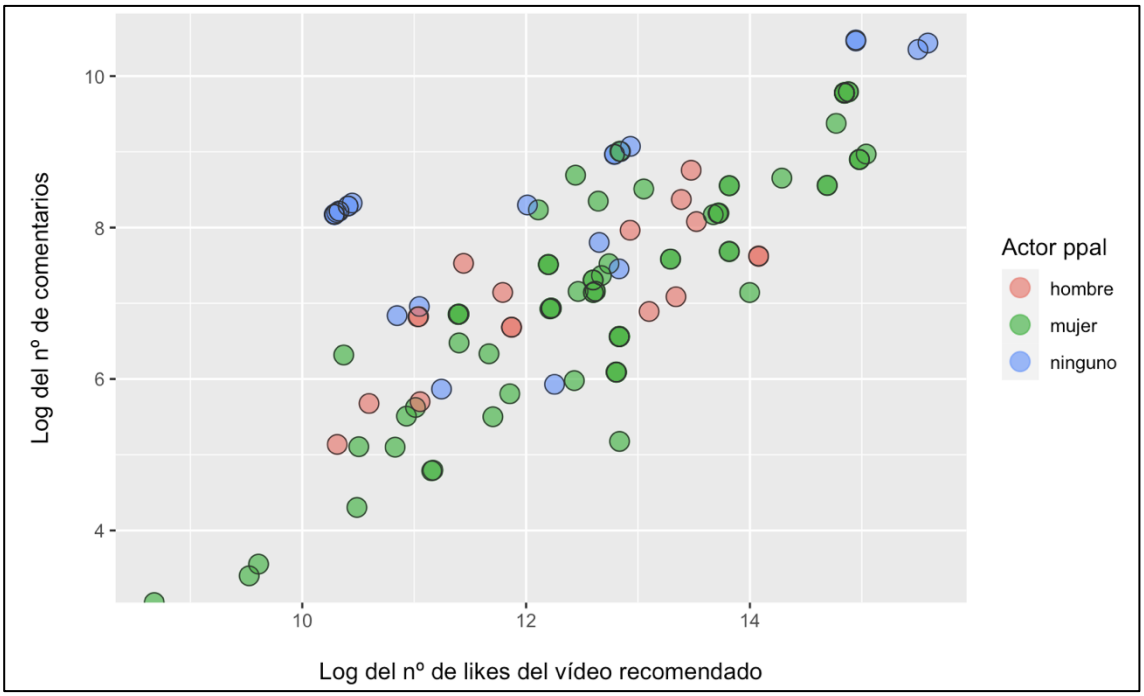


Fuente: Elaboración propia.

3.2.2. Entre las métricas de los vídeos recomendados.

Me gustas y comentarios del vídeo.

Figura 3.15: Diagrama de dispersión de los me gustas (log) y los comentarios (log).



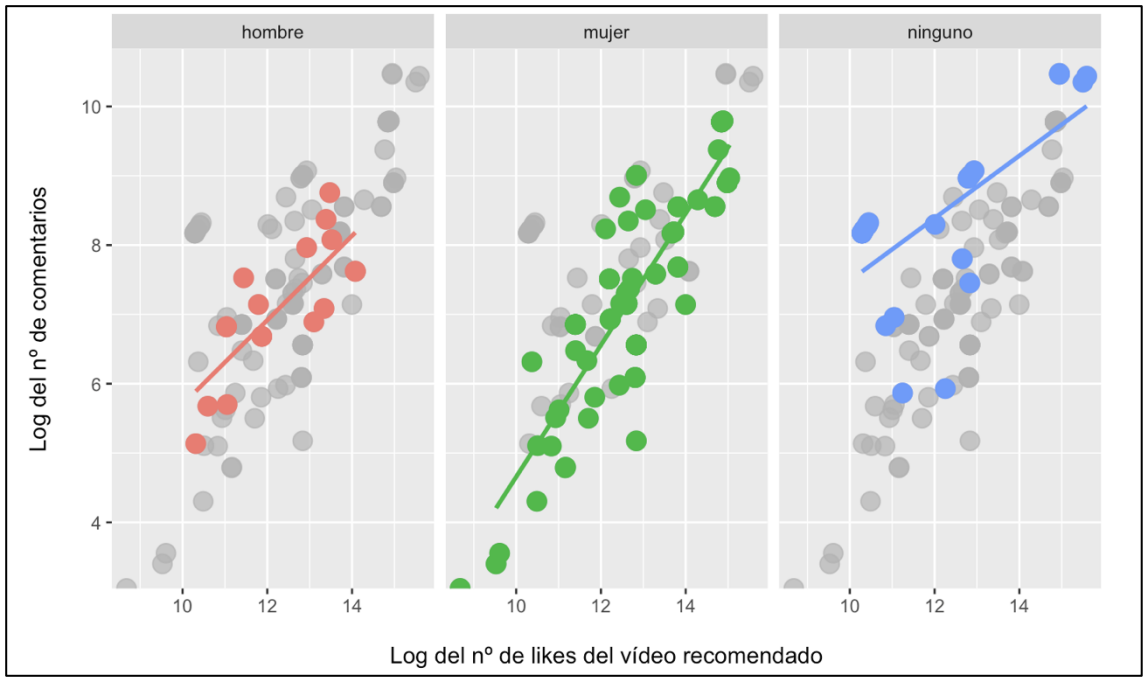
Fuente: Elaboración propia.

Al analizar la relación entre los comentarios y los me gustas de los vídeos recomendados, el rango de datos de ambas variables es tan amplio que es más interesante utilizar una escala logarítmica. El diagrama de dispersión muestra una clara correlación positiva. Vídeos con un gran número de *likes* son, generalmente, los que tienen un mayor número de comentarios, y viceversa.

Para cuantificar el grado de relación lineal que presentan estas variables, se ha calculado el coeficiente de correlación, r = 0,69. Este dato indica que la relación lineal que presentan estas variables es positiva, aunque moderadamente baja.

Si profundizamos en el estudio de esta relación segmentando por el actor principal del vídeo *Figura 3.16*, si bien las muestras son pequeñas, y se requiere precaución al extrapolar resultados, se aprecia que la relación no es tan clara cuando no se identifica ningún sexo como protagonista y la pendiente es mayor si el actor principal del vídeo es una mujer, es decir la relación es más pronunciada para este segmento.

Figura 3.16: Diagrama de dispersión de los me gustas (log) y los comentarios (log)en función del sexo del actor principal del vídeo.



Fuente: Elaboración propia.

Para estudiar si las diferencias entre los tres grupos son significativas se ha realizado un análisis de regresión. Los resultados de este análisis muestran que no existen diferencias significativas entre los tres niveles del factor considerados. La única variable que tiene un efecto significativo es la que mide el número de *likes*.

El actor principal del vídeo y el número de me gustas.

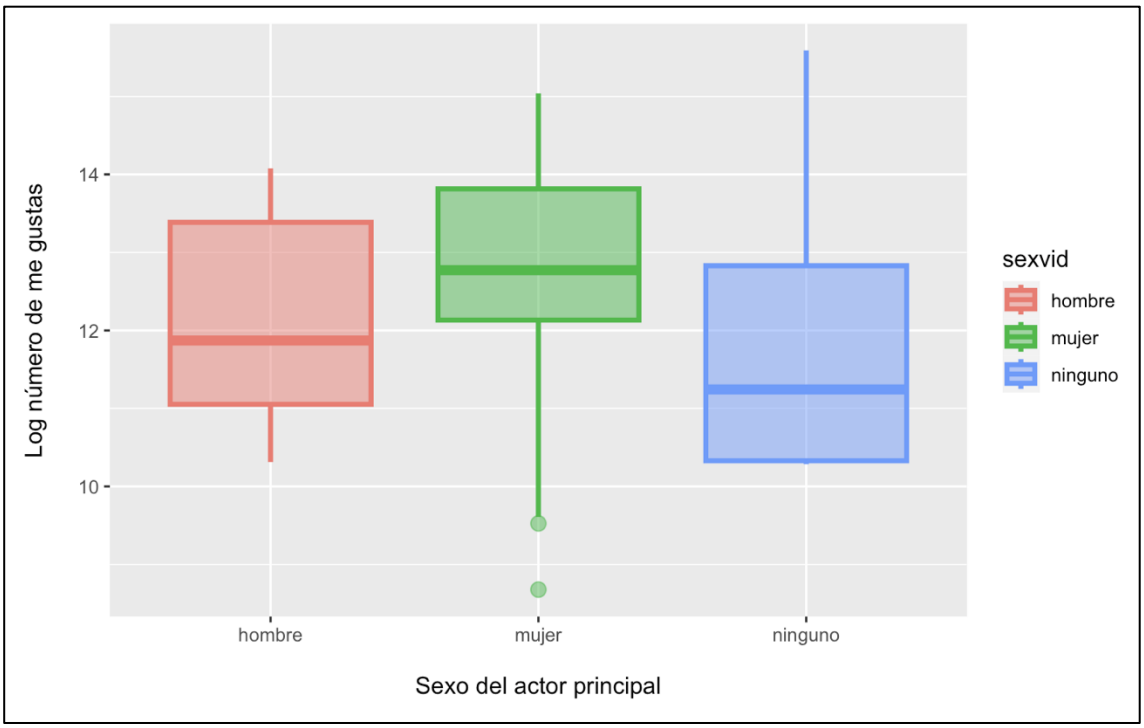
Para profundizar en este estudio se ha analizado la relación que tiene el sexo del actor principal con los me gustas que ha obtenido el vídeo. El gráfico de la figura 3.17 muestra diferencias entre los diferentes sexos. La primera diferencia que se aprecia es que las mujeres destacan cuando se trata de me gustas en vídeos que protagonizan. Cabe resaltar en este punto que, como se comentaba anteriormente, disponemos de más datos cuando la protagonista es una mujer, como se aprecia en el diagrama de

puntos. Por otra parte, cuando no se aprecia ningún sexo protagonista los me gustas son inferiores a cuando sí se identifica un hombre o una mujer.

Para analizar con más en detalle estas diferencias, los principales estadísticos del número de me gustas por grupos muestran que la mediana de las mujeres es superior a la de los otros grupos, llegando a duplicar a la de los hombres. En cuanto a los máximos y mínimos, las tres categorías de sexo se encuentran en rangos muy dispares entre sí.

Un Análisis de la Varianza, indica que las diferencias observadas sí son significativas (P=0.0436, obtenido mediante el test F del ANOVA, utilizando el logaritmo del número de *likes como variable respuesta*). El test de Tukey muestra que estas diferencias se dan entre las categorías de sexo: *ninguno* y *mujer*. No siendo significativas las diferencias entre hombres y mujeres.

Figura 3.17: Diagrama de cajas de los me gustas (log) según el sexo del actor principal del vídeo.



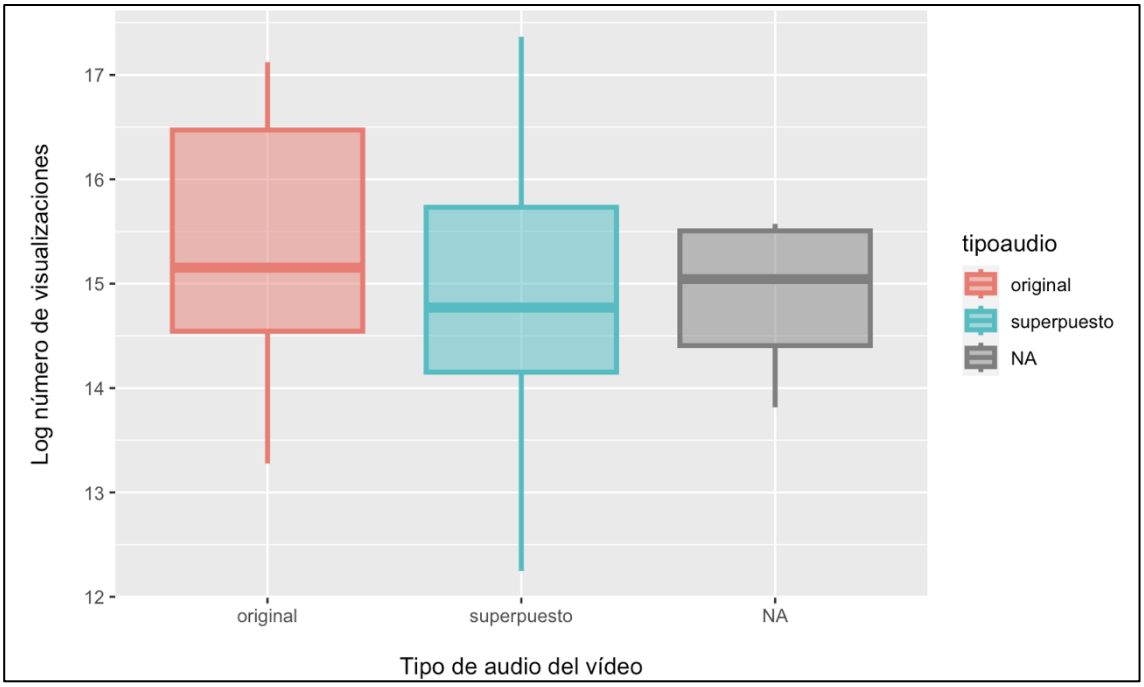
Fuente: Elaboración propia.

Tipo de audio y las visualizaciones del vídeo.

Como se comenta en el marco teórico, una parte importante para llegar a más visualizaciones es utilizar audios superpuestos de vídeos que son tendencia, para incrementar el número de *me gustas*. Hemos estudiado las diferencias en el número de visualizaciones de los vídeos recomendados según el tipo de audio que tenía el propio vídeo. En el diagrama de cajas se observan diferencias entre los tres tipos de audio.

Al realizar un análisis de la varianza se obtiene un valor P=0.0588 (obtenido mediante el test F del ANOVA y utilizando el logaritmo de las visualizaciones como variable respuesta). El análisis confirma, por tanto, que las diferencias observadas son prácticamente significativas para un nivel del 5%.

Figura 3.18: Diagrama de cajas de las visualizaciones (log) según el tipo de audio utilizado.

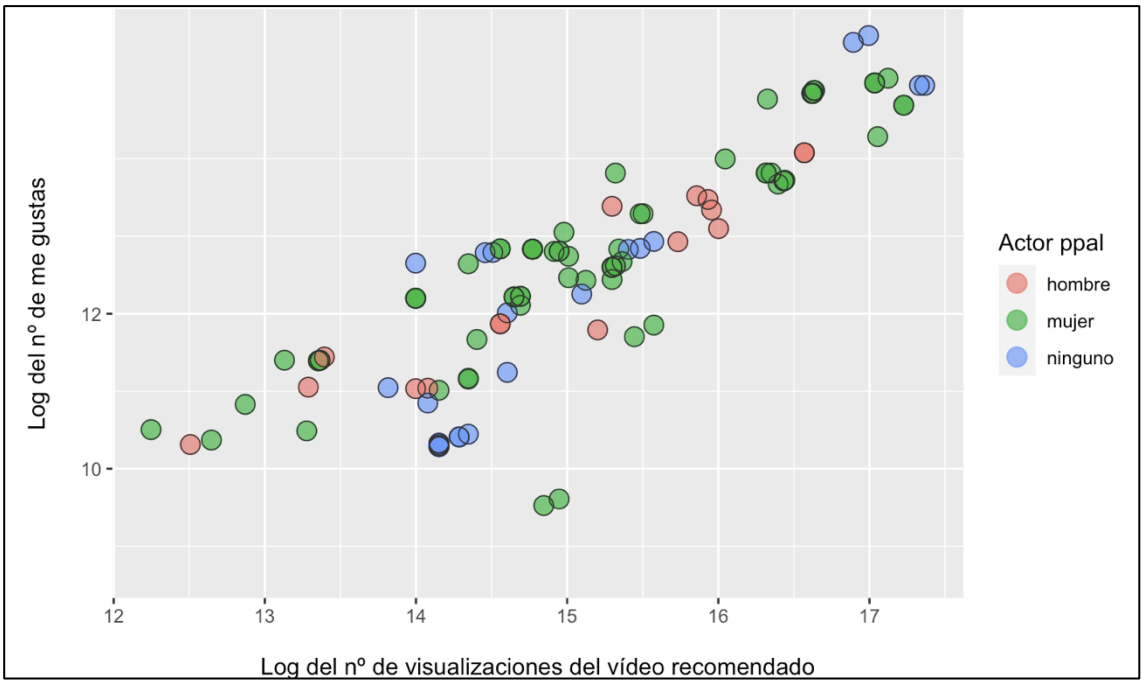


Fuente: Elaboración propia.

Visualizaciones y me gustas del vídeo.

Otro factor importante a la hora de observar las visualizaciones es la cantidad de me gustas que tiene el vídeo. Al tratarse de dos variables con un rango tan amplio de datos, para visualizar la posible relación es mejor realizar una transformación logarítmica.

Figura 3.19: Diagrama de dispersión de las visualizaciones (log) y el número de me gustas (log).

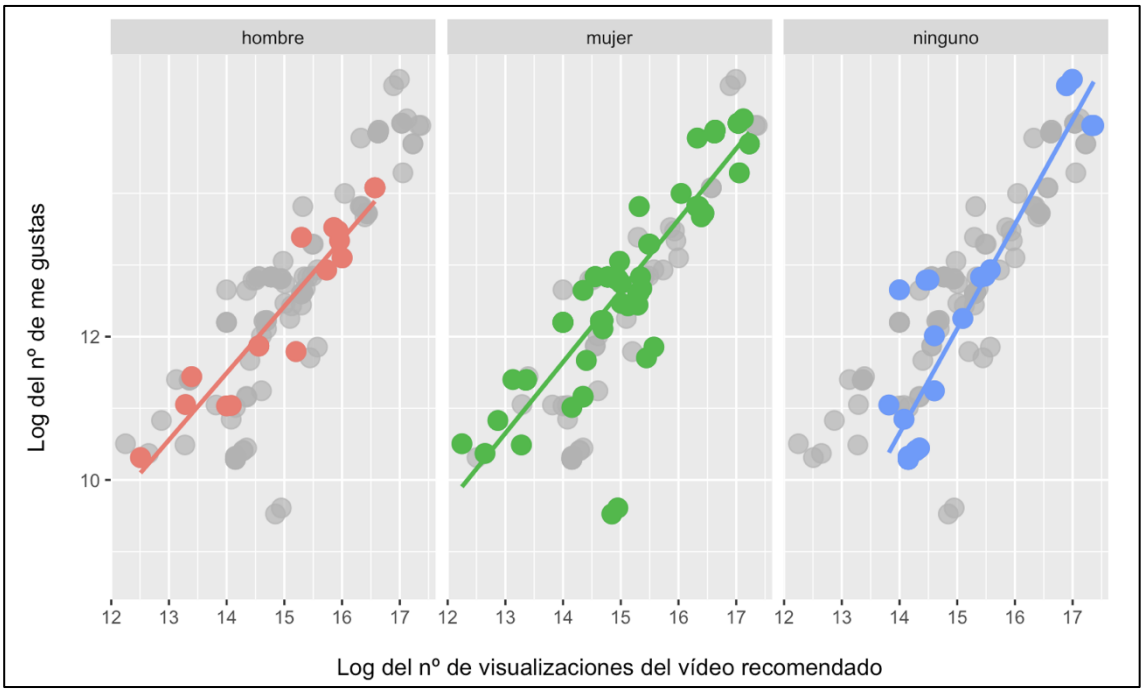


Fuente: Elaboración propia.

El diagrama de dispersión muestra una relación lineal positiva bastante clara, es decir, un incremento en el número visualizaciones tiene asociado, en general, un incremento en el número de *me gustas.*

Por otra parte, si se segmenta dicho diagrama por el sexo del actor principal de los vídeos no se percibe apenas diferencia entre ellos. Este dato quiere decir que, en principio, la relación entre las visualizaciones y los me gustas no se ve afectada por el sexo del protagonista del vídeo.

Figura 3.20: Diagrama de dispersión de las visualizaciones (log) y el número de me gustas (log) según el sexo del actor principal.



Fuente: Elaboración propia.

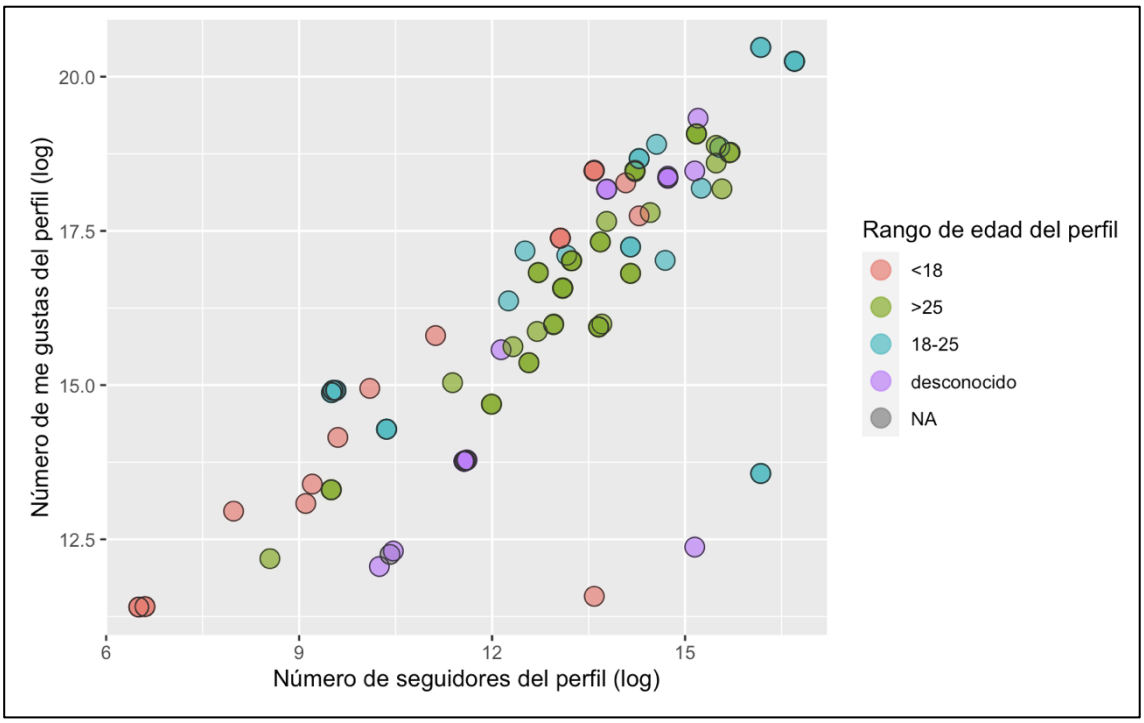
3.2.3. Entre las métricas de las cuentas que han generado los vídeos.

Seguidores y número de me gustas del perfil.

Una de las relaciones más fáciles de prejuzgar es que perfiles con mayor número de seguidores tendrán un mayor número de me gustas en su perfil. Para comprobar esta relación se ha realizado un diagrama de dispersión utilizando la transformación logarítmica en ambas variables debido a sus grandes varianzas. En la *Figura 3.21* se puede observar una clara relación lineal entre ambas variables.

Para determinar el grado de esta relación se ha calculado el coeficiente de correlación. En este caso, se está ante una relación positiva fuerte ya que estamos ante un coeficiente de 0.8116. Este dato nos lleva a decir que a mayor número de seguidores mayor número de *me gustas* tendrá el perfil y viceversa.

Figura 3.21: Diagrama de dispersión de los seguidores (log) y el número de me gustas del perfil (log) según el rango de edad del perfil.



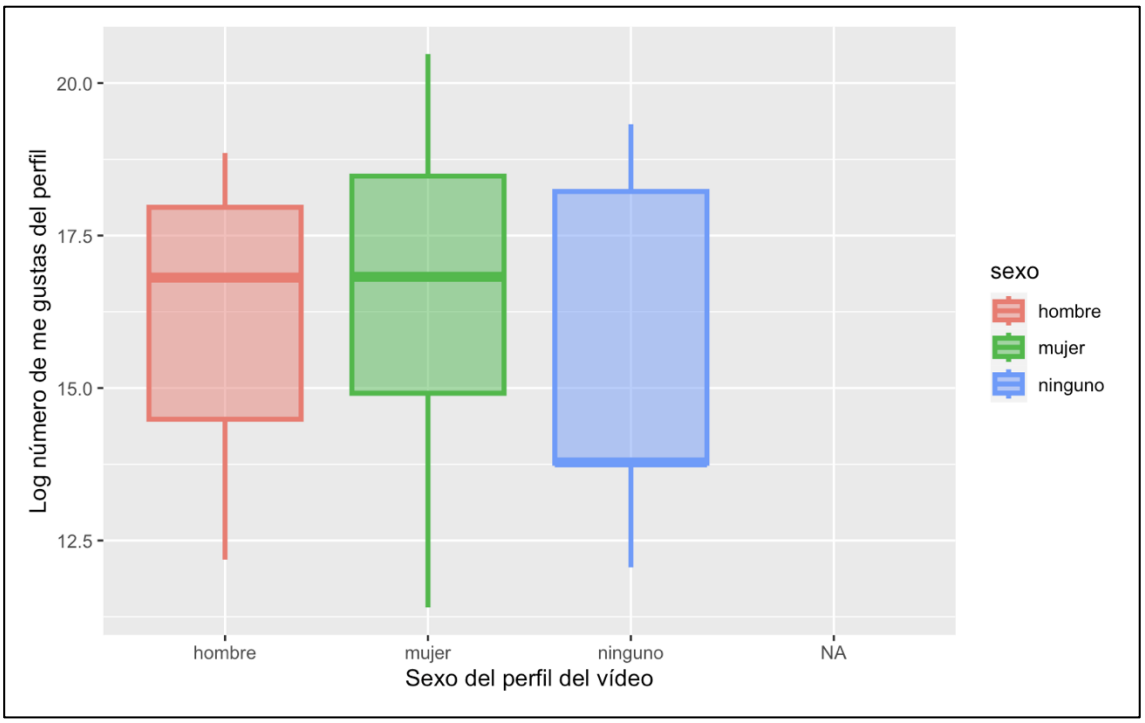
Fuente: Elaboración propia.

Sexo del perfil y el número de me gustas del perfil.

Otra relación interesante a considerar es ver si el sexo del perfil tiene algún tipo de relación con los *likes* totales del perfil. En la *Figura 3.22* se aprecian algunas diferencias: cuando no se identifica ningún sexo en el perfil la mediana de los me gustas es mucho más inferior a cuando sí se identifica uno de ellos. En cambio, entre los dos sexos las únicas diferencias destacables que se aprecian son entre los mínimos y los máximos ya que en las mujeres son mucho más extremos.

Al realizar el análisis de la varianza para comprobar si las diferencias son significativas, se ha obtenido un valor P=0.0605, por lo tanto, el sexo del perfil tiene prácticamente un efecto significativo sobre el número de *me gustas* de la cuenta.

Figura 3.22: Diagrama de cajas del número de me gustas del perfil (log) según el sexo del perfil.

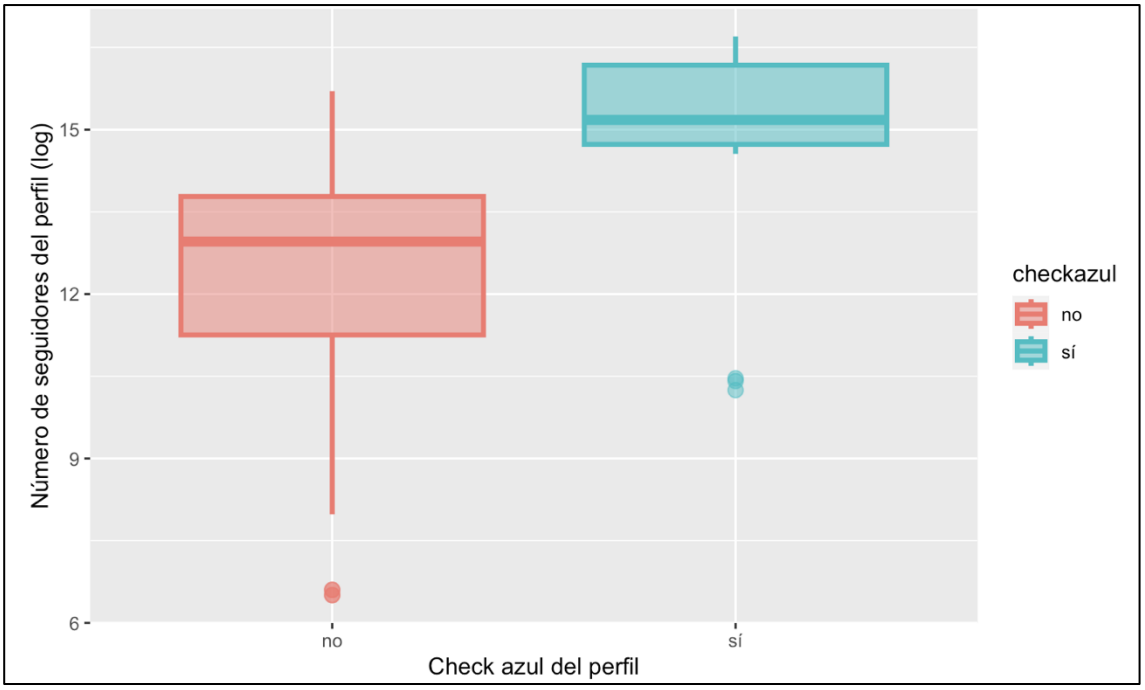


Fuente: Elaboración propia.

Check azul y el número de seguidores del perfil.

La verificación dentro de la plataforma de TikTok principalmente está presente con perfiles de gran cantidad de seguidores, por lo que es interesante analizar la relación que tienen estas variables dentro de los vídeos recabados.

Figura 3.23: Diagrama de cajas del número de seguidores del perfil (log) según el check azul.



Fuente: Elaboración propia.

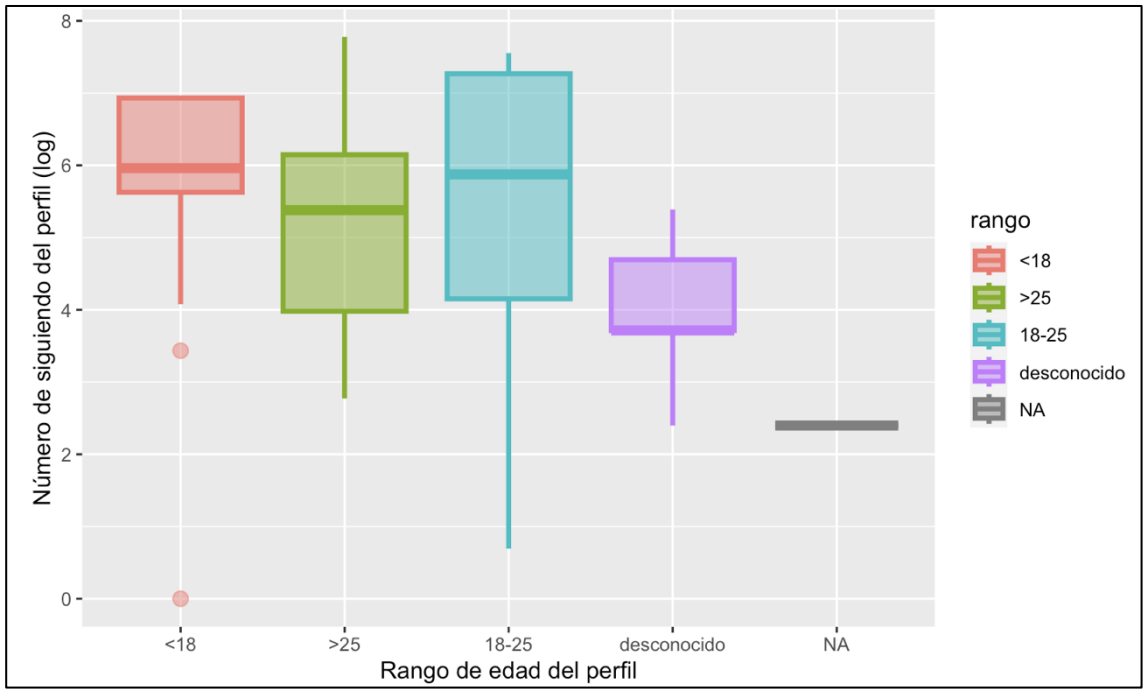
En la *Figura 3.23* se puede observar claramente cómo existe una gran diferencia entre los seguidores de los perfiles que sí tienen el *check azul* frente a los que no. Los perfiles que tienen la verificación se encuentran en un rango de seguidores mucho más elevado que para el resto de los perfiles, es decir, los que no tienen el *check azul*.

Un análisis de inferencia de comparación de medias proporcióna un valor P muy por debajo del 5%, lo que indica que las diferencias observadas son significativas.

El rango de edad y los siguiendo del perfil.

A la hora de analizar la relación del rango de edad del perfil y las personas que sigue se pueden observar algunas diferencias notables a primera vista.

Figura 3.24: Diagrama de cajas del número de siguiendo (log) según el rango de edad del perfil.



Fuente: Elaboración propia.

En la *Figura 3.24* se puede observar que cuando no se identifica un rango de edad dentro del perfil las métricas varían con respecto a cuando sí se identifica el rango de edad. La media del número de personas que sigue el perfil es mucho menor frente a las demás situaciones.

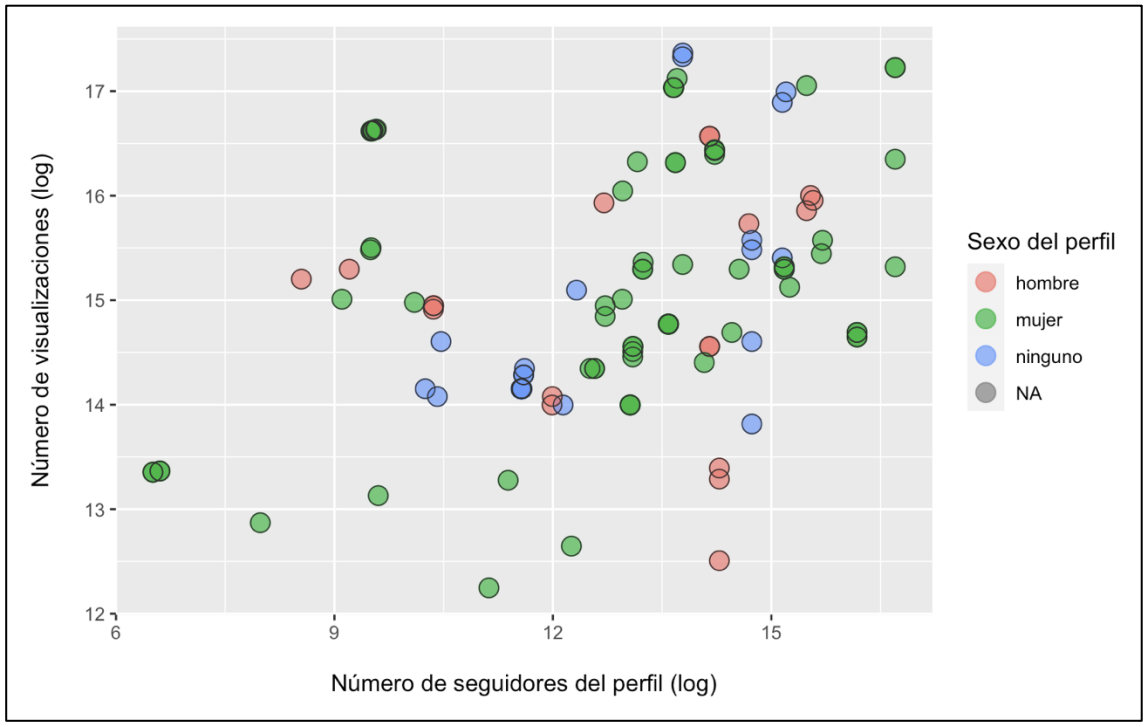
Al realizarse el análisis de la varianza (ANOVA) se obtiene un valor bastante bajo P=0.013 por lo que estamos ante una diferencia significativa. Este dato sugiere que el rango de edad del perfil influye en el número de siguiendo del propio perfil.

3.2.4. Entre las métricas de las cuentas y los vídeos recomendados.

Visualizaciones y el número de seguidores.

Las visualizaciones también podrían verse afectadas por el número de seguidores y, nuevamente, se realiza una transformación logarítmica debido a que las variables tienen rangos muy amplios en sus datos. A pesar de esto, no se encuentra ningún tipo de relación entre ambas variables en el diagrama de puntos. De hecho, el coeficiente de correlación calculado es de 0.306, lo que quiere decir que tienen una correlación positiva muy débil. Esto sugiere que el número de visualizaciones no está asociado de con el número de seguidores de un perfil. Este dato apoya lo que se comentaba en el marco teórico sobre la viralidad, ya que no necesariamente tener gran cantidad de seguidores genera más visualizaciones.

Figura 3.25: Diagrama de dispersión del logaritmo del nº de visualizaciones frente al logaritmo del número de seguidores, según el sexo del actor principal.



Fuente: Elaboración propia.

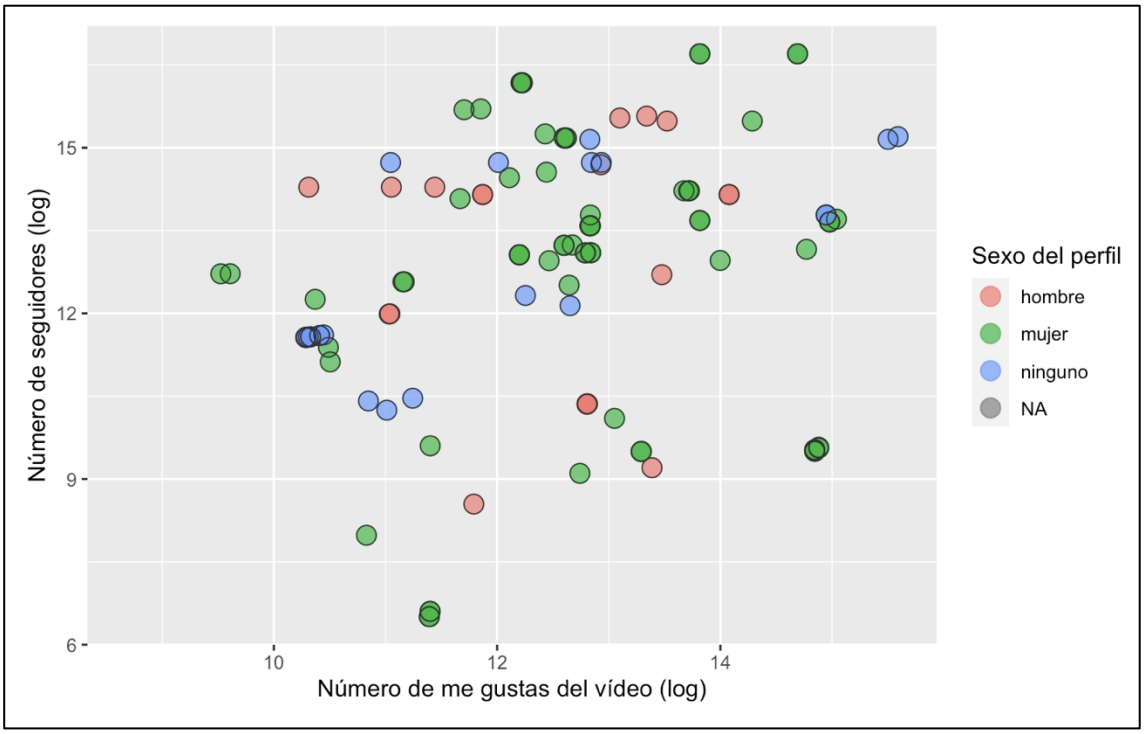
Me gustas del vídeo y seguidores del perfil.

Una de las principales características de TikTok que se comentaba en el marco teórico era la facilidad de conseguir viralidad sin necesariamente depender del número de seguidores. Debito a esto, es interesante saber la relación que existe en los datos recogidos sobre los *me gustas* de los vídeos y los seguidores de los perfiles.

Como primer contacto con la relación de las variables, en la *Figura 3.26* se puede observar que no se percibe ninguna relación clara. Esto sugiere que se podría estar cumpliendo la teoría comentada anteriormente.

El coeficiente de correlación que se obtiene para estas variables es del 0,19 por lo que, efectivamente, estamos ante una relación positiva muy débil. Esto quiere decir que no por el hecho de tener seguidores el vídeo subido tendrá más *me gustas*.

Figura 3.26: Diagrama de dispersión del logaritmo del nº de visualizaciones frente al logaritmo del número de seguidores, según el sexo del actor principal.



Fuente: Elaboración propia.

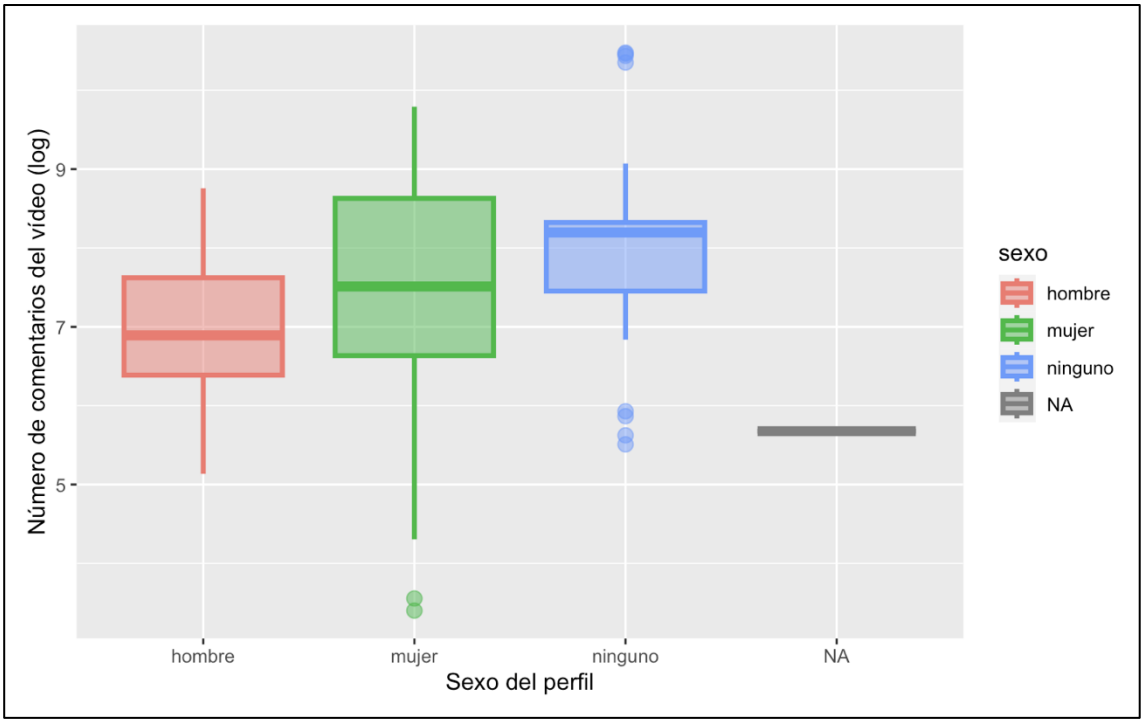
Sexo del perfil y comentarios del vídeo.

Otro aspecto importante es la relación que tiene el sexo del perfil en los comentarios que obtiene el vídeo. Dentro de la *Figura 3.27* se pueden observar varias diferencias entre las diferentes categorías del sexo de los perfiles.

Cuando no se identifica ninguno de los sexos se sitúan en un mayor rango y, por tanto, con una mayor mediana frente a las demás categorías. Cabe destacar que cuando el perfil pertenece a un hombre las métricas son menores que las de las mujeres.

Al realizar un análisis de la varianza (ANOVA) se obtiene un valor P=0.0439 por lo que las diferencias son significativas para un nivel del 5%. Esto quiere decir que a la hora de hacer comentarios en un vídeo recomendado, el sexo del perfil al que pertenece tiene un efecto significativo.

Figura 3.27: Diagrama de cajas del logaritmo del nº de comentarios el sexo del perfil.



Fuente: Elaboración propia.

4. MODELOS PREDICTIVOS.

4.1. Modelo de regresión logística.

Para la creación de un modelo predictivo hemos agrupado las categorías “baile”, “cocina y humor” y “noticias en tendencia” en la categoría de “otro” para evitar las categorías con un número reducido de datos. Además, se ha creado una nueva variable a través del nombre de los perfiles en la que se recoge si un perfil (cuenta de TikTok) ha recomendado, o no, más de un vídeo.

El modelo de regresión logística utilizado pretende predecir si una cuenta de TikTok ha recomendado más de un vídeo, pudiendo así valorar si se puede predecir que un perfil se repita en función de otras variables. Tras varias pruebas de modelos se han ido descartando aquellos en los que no aparecían variables significativas.

El modelo final es el que se puede observar en la *Figura 3.28*, en el que se encuentran como variables explicativas el sexo del perfil, el número de personas que sigue el perfil del vídeo recomendado y la interacción entre ambas. En este modelo la interacción entre el sexo y los siguiendo son significativos si consideramos un nivel de significación del 10%.

Figura 3.28: Modelo de regresión logística para la variable respuesta “ctarec”.

```
Call:
glm(formula = ctarec ~ sexo + siguiendo + sexo:siguiendo, family = binomial,
    data = ctas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8199  -0.9885   0.1584   1.0575   1.4591

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.261449   1.113468   1.133   0.2573
sexomujer             -1.787035   1.224326  -1.460   0.1444
sexoninguno            1.444905   2.071502   0.698   0.4855
siguiendo             -0.007463   0.005104  -1.462   0.1437
sexomujer:siguiendo    0.009514   0.005277   1.803   0.0714 .
sexoninguno:siguiendo -0.015910   0.016519  -0.963   0.3355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.455  on 52  degrees of freedom
Residual deviance: 59.255  on 47  degrees of freedom
AIC: 71.255

Number of Fisher Scoring iterations: 6
```

Fuente: Elaboración propia.

El modelo de regresión consigue clasificar correctamente el 64.15% de las cuentas. A primera vista, parece que el modelo de regresión logística funciona mejor

que la adivinación aleatoria. Sin embargo, este resultado es engañoso porque hemos entrenado y probado el modelo con el mismo conjunto de 53 observaciones. En otras palabras, 100% - 64.15% = 35.85%, es la tasa de error para la BBDD utilizada para ajustar el modelo.

Esta tasa suele ser optimista, y tiende a subestimar la tasa de error para una base de datos de prueba. Para evaluar mejor la precisión del modelo de regresión logística, deberíamos examinar lo bien que predice nuestro modelo utilizando una BBDD distinta. De este modo, obtendríamos una estimación de la tasa de error del modelo más realista.

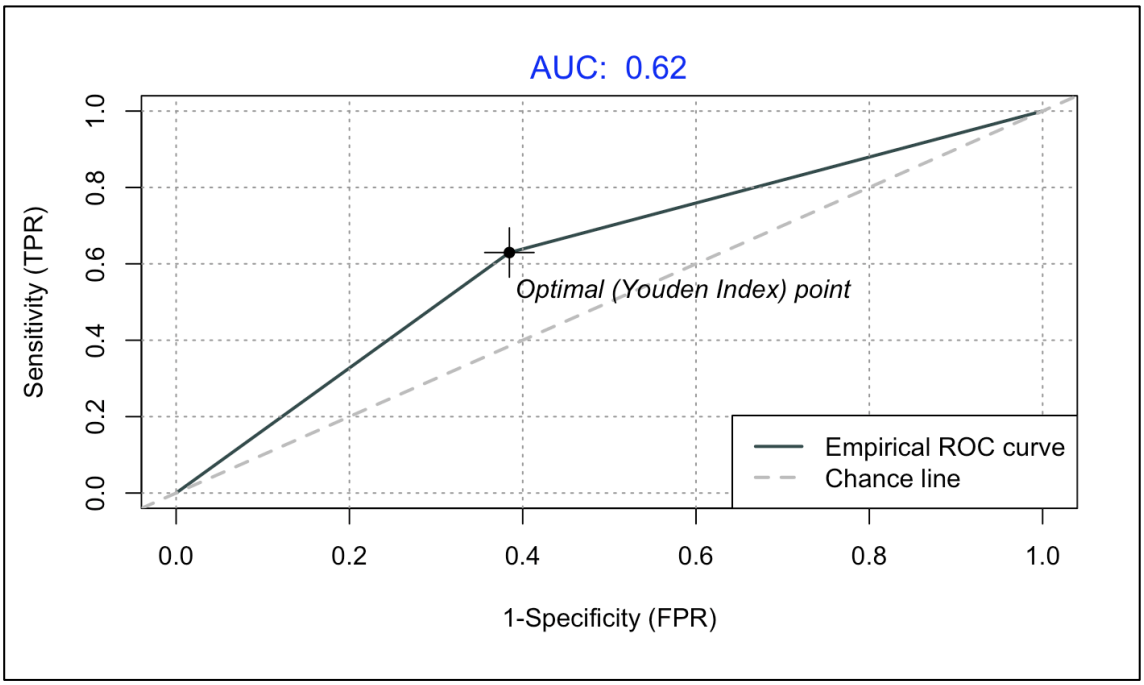
Como la base de datos con la que trabajamos es pequeña (cuenta sólo con 99 casos), en lugar de dividirla en dos conjuntos de datos (entrenamiento y test), optamos por estimar la tasa de error a partir de la técnica de Validación Cruzada.

Mediante validación cruzada la estimación de la tasa de error se ve incrementada ligeramente a un 37.74%, por lo que la precisión total del modelo es del 62%.

Mediante la curva ROC y el análisis del índice de Youden se puede encontrar el valor de corte óptimo, es decir, el valor que proporciona el mejor equilibrio entre sensibilidad y especificidad.

La sensibilidad nos indica la capacidad de nuestro modelo de clasificar como casos positivos las cuentas que realmente sí han proporcionado más de un vídeo. La especificidad nos indica la capacidad de nuestro modelo de clasificar como casos negativos las cuentas que realmente no han proporcionado más de un vídeo.

Figura 3.29: Curva ROC para las predicciones obtenidas mediante LOOCV en el modelo de regresión logística.



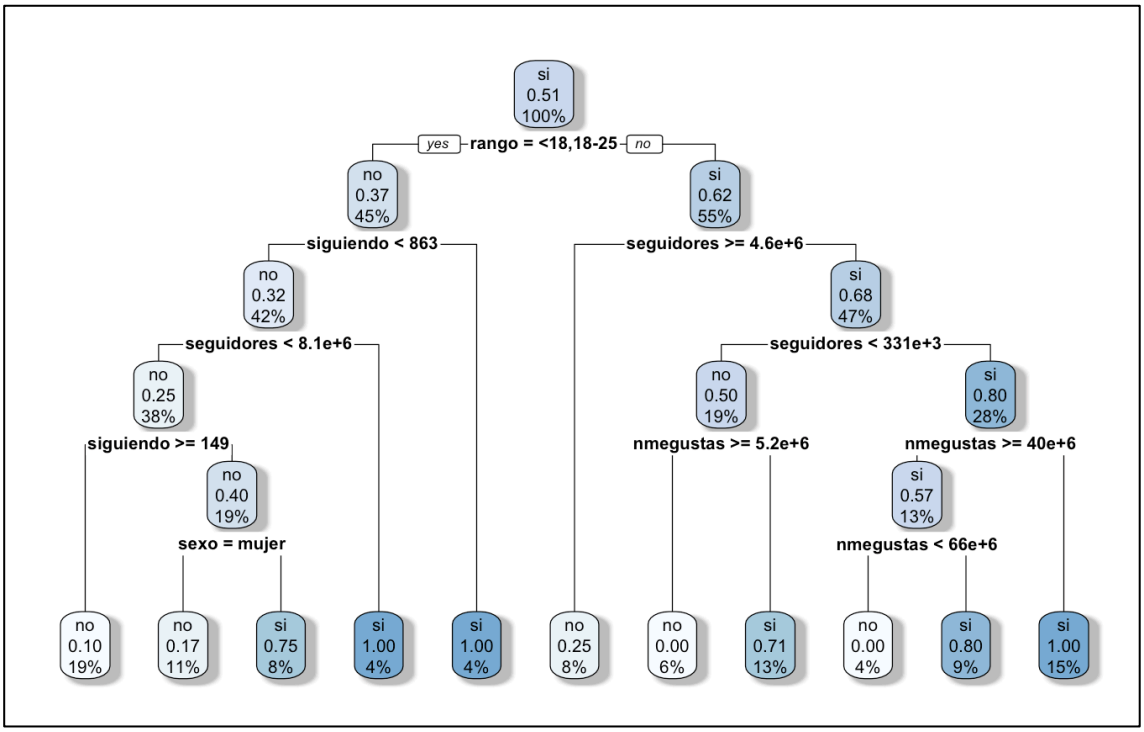
Fuente: Elaboración propia.

4.2. Árbol de decisión.

Por otra parte, se ha creado un árbol de decisión con el objetivo de probar varios modelos predictivos y valorar cuál de ellos predice mejor los valores de nuestra variable respuesta.

El árbol de decisión creado está diseñado para predecir si una cuenta recomienda o no más de un vídeo en función de las demás variables. Como en el modelo de regresión logística, se predice la variable creada en el modelo anterior "ctarec".

Figura 3.30: Árbol de decisión predicción si una cuenta recomienda o no más de un vídeo.



Fuente: Elaboración propia.

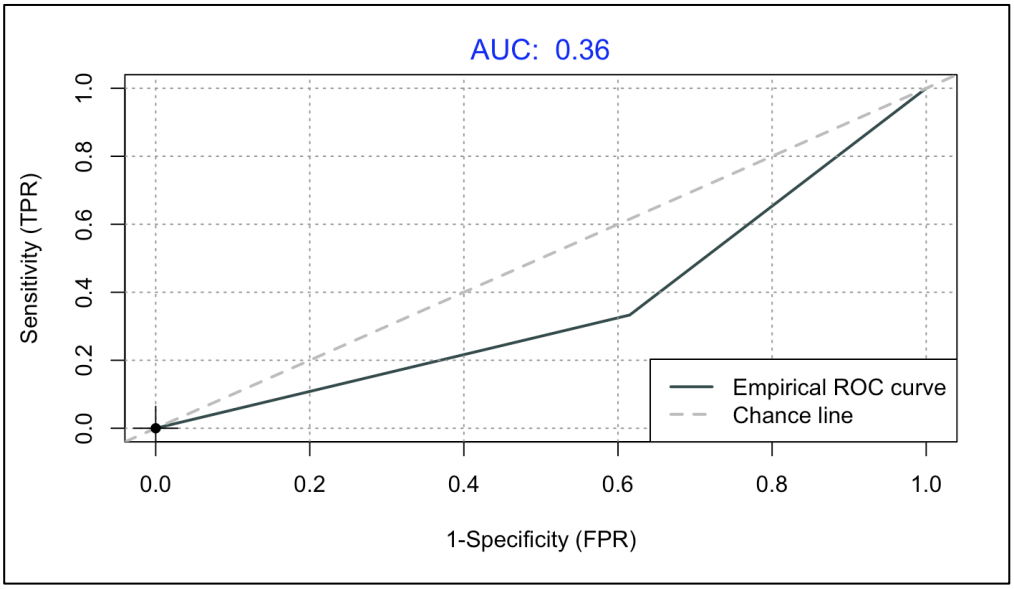
En la *Figura 3.30* se pueden observar las variables que más peso tienen en la clasificación de la predicción son el rango de edad del perfil, el número de personas que sigue y los seguidores que tiene el perfil. Cabe destacar que todas ellas son parte de las variables que hacen referencia al perfil de los vídeos recomendados.

La tasa de error del árbol, calculada con la misma base de datos que hemos utilizado para construirlo, es del 13%. No obstante, como comentamos en el modelo anterior, este dato suele sobre-estimar la tasa de error ya que podría ser muy diferente a la que obtendríamos si se utilizara otra base de datos distinta a la utilizada para construir el árbol de decisión. Como no se dispone de suficientes datos para dividir la BBDD en dos conjuntos, uno para ajustar el modelo y otro para validarlo, se opta por utilizar la técnica de Validación Cruzada.

A través de la Validación Cruzada se obtiene una tasa de error del 64%, en este caso aumenta considerablemente frente a la anterior.

En la *Figura 3.31* se puede observar el bajo nivel de predicción que tiene el modelo. La curva ROC está por debajo del 50% de precisión, lo que indica que este modelo funciona incluso peor que la utilización del simple azar.

Figura 3.31: Curva ROC para las predicciones obtenidas mediante LOOCV en el árbol de decisión.



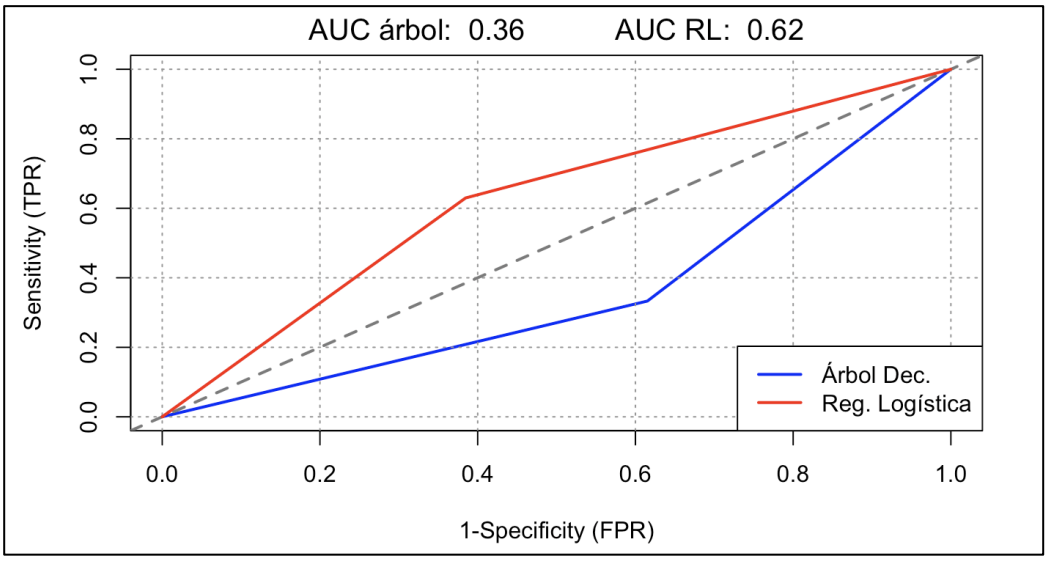
Fuente: Elaboración propia.

El árbol de decisión es mucho peor que el modelo de regresión logística que se ha utilizado anteriormente. El árbol de decisión no llega ni siquiera a una precisión del 50% para predecir si una cuenta recomienda más de un vídeo.

4.3. Comparación modelos.

Para comparar ambos modelos de una manera visual se pueden representar las curvas ROC de ambos modelos en un mismo gráfico.

Figura 3.32: Curva ROC para las predicciones obtenidas mediante LOOCV en el árbol de decisión.



Fuente: Elaboración propia.

El valor máximo del índice de Youden es 1 (ajuste perfecto) y el mínimo es 0 (el modelo no tiene capacidad predictiva). El mínimo se obtiene cuando sensibilidad = 1 - especificidad, representado por la diagonal principal del gráfico de la curva ROC. La distancia vertical entre la diagonal y la curva ROC en un punto de corte concreto, proporciona el valor del índice para ese punto.

El resultado ideal se obtendrá cuando la sensibilidad es igual a 1 y al mismo tiempo la especificidad es 1, es decir, cuando la tasa de falsos positivos (1-especificidad) es 0. Cuanto más se acerque una curva ROC a esta situación ideal, mejor será el rendimiento del modelo si consideramos que la sensibilidad y la especificidad tienen la misma importancia. El punto que representa esta combinación estará en la esquina superior izquierda del gráfico.

Teniendo en cuenta esto, se puede observar en la *Figura 3.32* claramente que el modelo de regresión logística es un modelo mucho mejor que el modelo basado en un árbol de decisión. Este último es peor que el de regresión e incluso peor que la simple asignación aleatoria de las cuentas en las dos clases consideradas, ya que el área bajo la curva ROC para este modelo es mucho menor, y menor del 50%.

5. CONCLUSIONES

TikTok comenzó su auge cuando la pandemia mundial hizo que las personas se vieran obligadas a quedarse en casa y tener que buscar nuevas vías de entretenimiento. No obstante, aunque este haya sido su comienzo, ha sabido afianzarse como una de las redes sociales más usadas a nivel mundial, llegando a ser la primera plataforma con una presencia de estas magnitudes creada en China.

No solo destaca como una de las plataformas sociales favoritas, sino que ha impuesto la tendencia de los vídeos cortos al resto de sus competidores. La facilidad de consumo de este tipo de contenido es lo que lo hace ser un esencial en las redes sociales hoy en día.

No se puede hablar del éxito de TikTok sin destacar su característica forma de mostrar el contenido gracias a su algoritmo. Utilizando la inteligencia artificial es capaz de mostrar contenido a los usuarios en función de las interacciones a tiempo real que se realizan mientras los usuarios pasan tiempo dentro de la plataforma.

Actualmente TikTok es la red social de la generación Z por excelencia y así lo muestra su audiencia, ya que gran parte de la misma se ubica en esta generación. Esto hace que sea una audiencia más exigente con el contenido que visualiza y destaque el *fast content* con mensajes sencillos y claros.

El gran auge de la red social y el aumento del gasto en publicidad en redes sociales ha hecho que TikTok haya tenido que adaptarse y crear una plataforma propia para las marcas denominada TikTok for Business.

La plataforma también tiene una parte negativa ya que es difícil comprobar la veracidad de la información que difunde debido a la velocidad de difusión de la misma. Este problema es más serio cuando abarca temas como la salud y el bienestar, o la igualdad de género, por las consecuencias de la difusión de bulos asociados con estos temas.

El estudio llevado a cabo, asumiendo las limitaciones en cuanto al tamaño de la muestra y a que la geolocalización y la fecha en la que se ha realizdo el estudio afectan claramente a los resultados, muestra los factores que repercuten en las primeras recomendaciones de TikTok en cuentas recién registradas.

Como era de esperar según el marco teórico, la temática más presente en los vídeos recabados es el humor. Además, en su gran mayoría el actor principal del vídeo recomendado e una mujer.

Dentro de las cuentas de los vídeos, se pudo observar que la mayoría de ellos pertenecían a personas con un rango de edad de mayores de 25 y entre los 18 y 25 años. Perfiles que pertenecen principalmente a la generación Z. Además, la mayor parte de los perfiles no tenían el verificado de TikTok con el *check azul*.

La relación entre *me gustas* y comentarios se ha verificado que se trata de una correlación positiva y moderadamente alta, a más *me gustas* más comentarios y viceversa. Sucede lo mismo con las visualizaciones y los me gustas, también existe una relación positiva entre ambas variables. Una relación significativa que llama la atención es la que existe entre el actor principal y los *me gustas*: cuando el actor principal es una mujer, frente a cuando no se identifica ningún sexo, el vídeo recomendado obtiene más *me gustas*.

Comparando las métricas de los perfiles se han obtenido varias diferencias significativas. Una de ellas es que los seguidores y el número de *me gustas* total del perfil tienen una relación positiva fuerte, lo que significa que a mayor número de seguidores mayor número de *me gustas* totales tendrá el perfil. La otra diferencia es la relación que existe entre los seguidores y el *check azul*.

Se ha estudiado también la relación entre las visualizaciones y los seguidores. En este caso, su relación es muy baja, lo que apoya el discurso de que en TikTok no hace falta tener seguidores para que tu vídeo se muestre a los usuarios. Otro argumento que apoya esta afirmación es que los me gustas están muy poco relacionados con los seguidores: no necesariamente necesitas tener un gran número de seguidores para alcanzar buenas cifras de me gustas.

Por otra parte, entre el sexo del perfil y los comentarios se encuentra una diferencia significativa. Lo que quiere decir que el sexo influye en la predisposición de los usuarios a comentar los vídeos.

También se han obtenido conclusiones de los dos modelos predictivos que se han diseñado para intentar predecir si una cuenta realiza más de una recomendación o no. Se ha valorado que es más efectivo el modelo de regresión logística en el que las variables que se toman como variables explicativas son el sexo del perfil, el número de cuentas que sigue, y la interacción entre ambas variables.

En definitiva, TikTok destaca por su exitosa forma de mostrar vídeos, su facilidad de uso y su capacidad de hacer viral a cualquier usuario, centrándose en el contenido. Las conclusiones de este estudio deben valorarse asumiendo las limitaciones del mismo, ya que la geolocalización y la fecha del estudio delimitan la población objetivo al que pueden extrapolarse los resultados.

6. BIBLIOGRAFÍA

Alonso López, N., Giacomelli, F. y Sidorenko Bautista, P. (2021). Espacios de verificación en TikTok. Comunicación y formas narrativas para combatir la desinformación. *Revista Latina de Comunicación Social*, 79, 87-113.

Bahiyah Omar, W. (2020). Watch, Share or Create: The influence of personality traits and user motivation on TikTok mobile video usage*. International Journal of Interactive Mobile Technologies*, 14 (4), 121-137.

Ballesteros Herencia, C. A. (2020). La propagación digital del coronavirus: Midiendo el engagement del entretenimiento en la red social emergente TikTok. *Revista española de comunicación en salud*, suplemento 1, 171-185.

Benavides, C., Feijoo, B. y Pavez, I. (2023). Análisis de las percepciones de los menores chilenos sobre el contenido comercial en TikTok: "Me comí un anuncio". *Revista Mediterránea de Comunicación*, 14 (2), 1-10.

Elhai, J. D., Montag, D. y Yang, H. (2021). On the psychology of TikTok use: a first glimpse from empirical findings. *Frontiers in Public Health*, 9, 1-6.

Cătălina – Oana, F. (2021). Motivation of TikTok users. *International Journal of Current Science Research and Review*, 4 (12), 1640-1644.

Camacho Gómez, A. S., Cardozo Morales, Y. y Cristancho Triana, G. J. (2022). Tipo de centennials en la red social TikTok y su percepción hacia la publicidad. Revista CEA, 8 (17), e1933.

Conde del Rio, M. A. (2021). Estructura mediática de TikTok: estudio de caso de la red social de los más jóvenes. *Revista de Ciencias de la Comunicación e Información*, 26, 59-77.

García Marín, D. y Salvat Matinrey, G. (2022). Viralizar la verdad. Factores predictivos del engagement en el contenido verificado en TikTok. Profesional de la información, 31 (2), e310210.

García Jiménez, A. y Suárez Álvarez, R. (2021). Centenials en TikTok: tipología de vídeos. Análisis y comparativa España-Gran Bretaña por género, edad y nacionalidad. *Revista Latina de Comunicación Social*, 79, 1-22.

Gareth J. et al. (2021) An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics. Second Edition.

Herranza de la Casa, J. M., Moya Ruiz, A. S. y Sidorenko Bautista, P. (2021). Análisis de la comunicación de empresas europeas y norteamericanas en TikTok. *AdResearch ESIC International Journal of Communication Research*, 25 (25), 106-123.

Kallner, Anders. (2018) Laboratory Statistics. Methods in Chemistry and Health Sciences. *Elsevier*. Second Edition.

Larrondo-Ureta, A., Morales-i-Gras, J. y Peña-Fernández, S. (2022). Información de actualidad en TikTok. Viralidad y entretenimiento para nativos digitales. *Profesional de la información*, 31 (1), 1-13.

Maldonado López, S. B., Oswaldo Dután, W. Y Romero Quiroga, K. R. (2023). Creación de contenidos en el nuevo esquema de comunicación masiva. *Revista Científica Mundo de la Investigación y el Conocimiento*, 7 (1), 398-406.

Martín Ramallal, P., Merino Cajaraville, A. y Micaletto Belda, J. P. (2022). Contenidos digitales en la era de TikTok: percepción de los usuarios del botón Covid-19 en España. *Revista de Comunicación y Salud*, 1, 1-23.

Martín Ramallal, P. y Micaletto Belda, J. P. (2021). Tiktok, red simbiótica de la generación z para realidad aumentada y el advergaming inmersivo. *Revista de Comunicación*, 20 (2), 223-242.

Méndez Majuelos, M. I. y Olivares-García, F. J. (2020). Análisis de las principales tendencias aparecidas en TikTok durante el periodo de cuarentena por la COVID-19. *Revista española de comunicación en salud*, suplemento 1, 243-252.

R Core Team (2023). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.

ANEXO I. RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030.

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**			X	
ODS 4. **Educación de calidad.**		X		
ODS 5. **Igualdad de género.**		X		
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades**.				X
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**	X			
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**			X	
ODS 17. **Alianzas para lograr objetivos.**			X	

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

***Utilice tantas páginas como sea necesario.

Editor: Secretaria General / UPV · D.L.: V-5092-2006 · ISSN: 1887-2298 · www.upv.es/secgen

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030. (Numere la página)

Este trabajo tiene la relación que se describe a continuación con los siguientes ODS:

- ODS 3 - Salud y bienestar: TikTok puede ser una espacio para difundir información relevante sobre la salud pública. Además, puede ser una gran vía de comunicación para crear concienciación sobre temas de vida saludable.

- ODS 4 - Educación de calidad: la red social puede llegar a ser una nueva vía para tratar información educativa de una manera innovadora y más creativa. Esta vía puede ser una opción más interactiva para la educación.

- ODS 5 - Igualdad de género: con la gran visibilidad que aporta TikTok se puede utilizar para romper estereotipos, dar voz a discursos de igualdad y que no se encuentre ningún tipo de barrera.

- ODS 9 - Industria, innovación e infraestructuras: como plataforma relativamente nueva TikTok forma parte de la innovación y además también invita a la creatividad de sus usuarios.

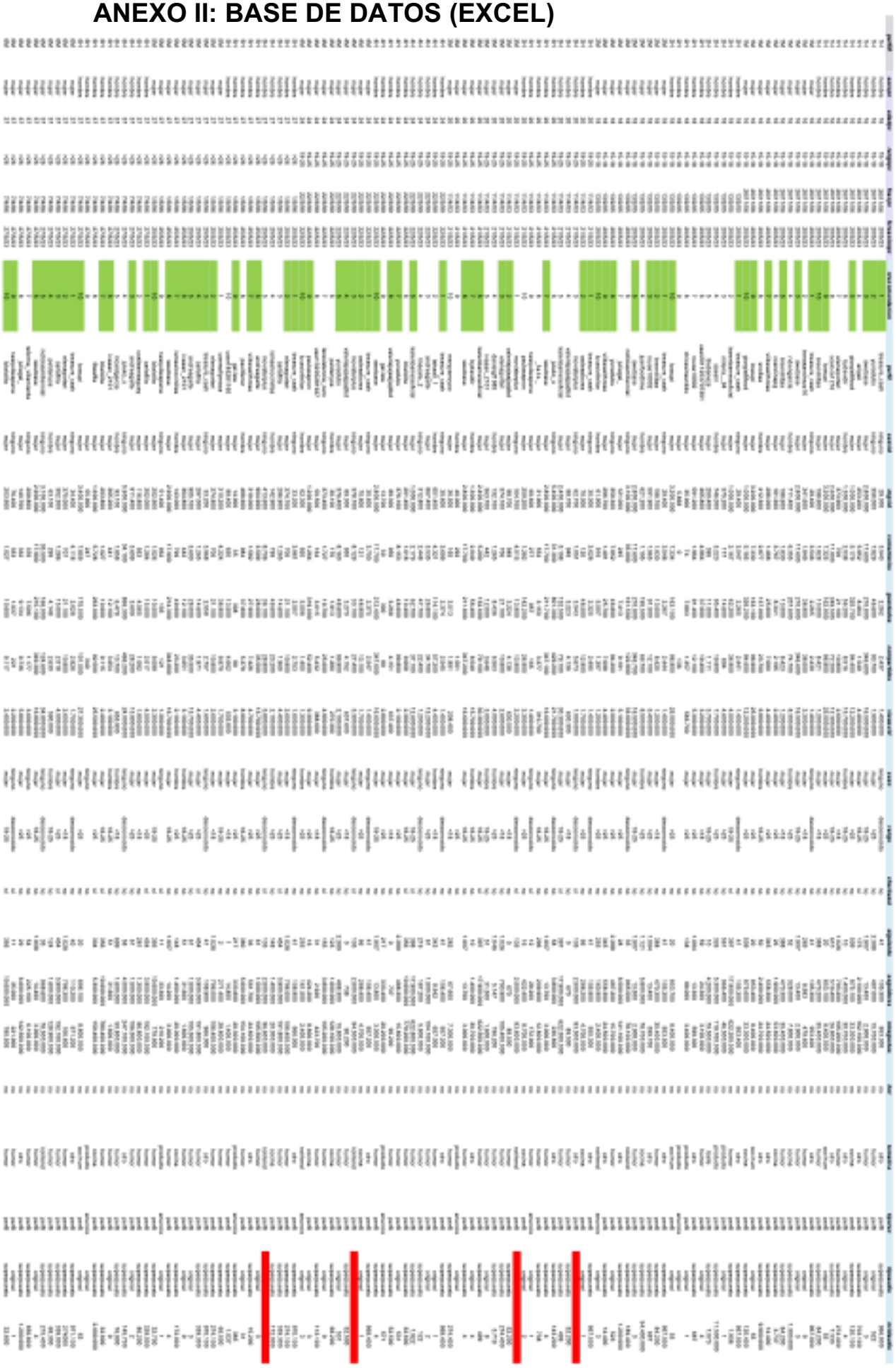
- ODS 13 - Acción por el clima: como red social, TikTok es un gran canal de difusión para visibilizar y dar conciencia del medioambiente. A través de la plataforma se puede dar voz a pronblemas climáticos que están ocurriendo y también a concienciar sobre hábitos sostenibles.

- ODS 16 - Paz, justicia e instituciones sólidas: TikTok puede llegar a ser un espacio donde se intente sensibilizar sobre los derechos humanos, la justicia social y el orden pacífico.

- ODS 17 - Alianzas para lograr objetivos: la red social puede coloborar en la generación de campañas por parte de las ONGs y difundir la concienciación de los propios ODS.

Editor: Secretaria General / UPV · D.L.: V-5092-2006 · ISSN: 1887-2298 · www.upv.es/secgen

ANEXO II: BASE DE DATOS (EXCEL)



ANEXO III: ARCHIVO R STUDIO

Base de datos *tiktok.xlsx*

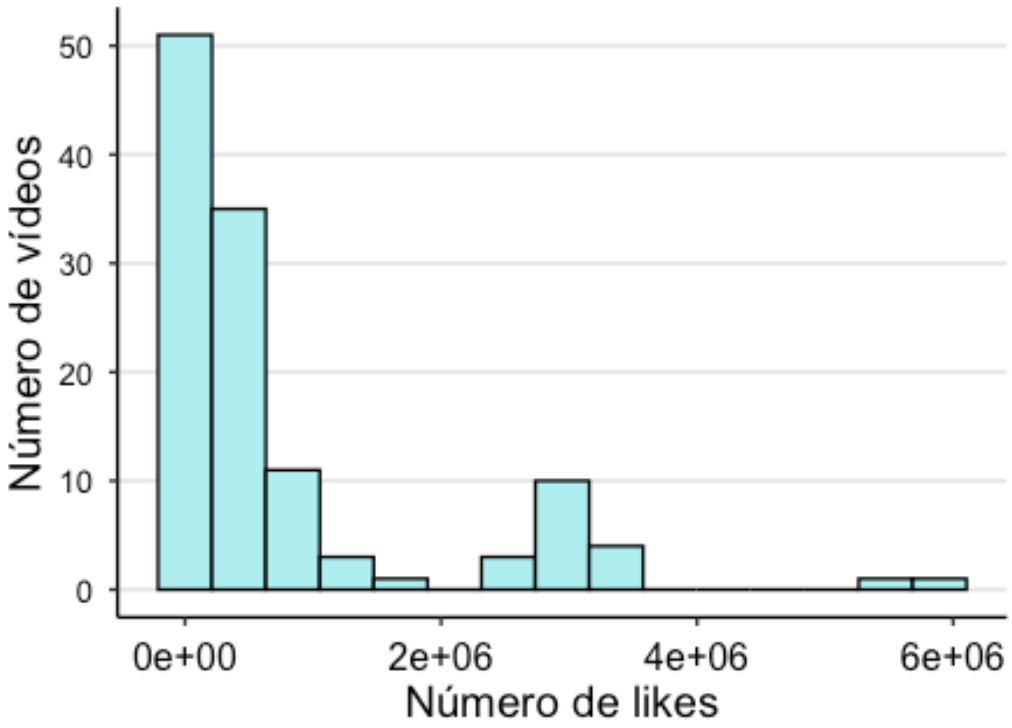
3. RESULTADOS DEL ANÁLISIS EXPLORATORIO

3.1. Análisis descriptivo de las variables

3.1.1. Métricas de los vídeos recomendados

Me gustas de los vídeos recomendados.

```
ggplot(data=tiktok, aes(x=mgvid)) +
  geom_histogram(bins = 15, fill="paleturquoise", color="black") +
  labs(title=" ",
    x="Número de likes",
    y="Número de vídeos")
```

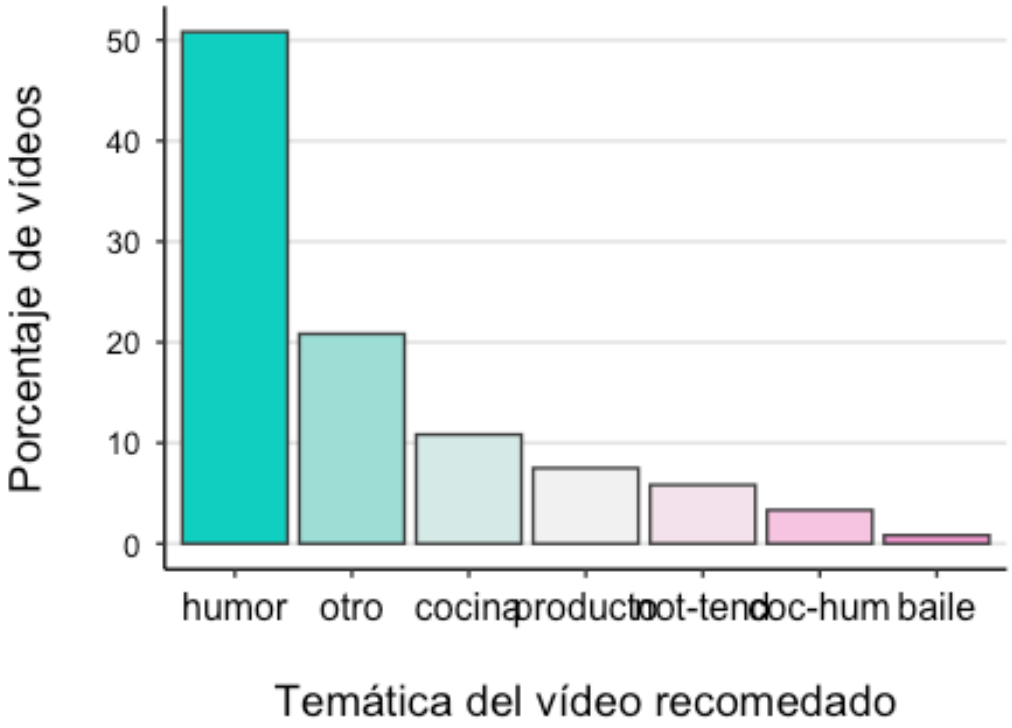



```
summary(tiktok$mgvid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5868   88700  306250  775715  876775 5900000
```

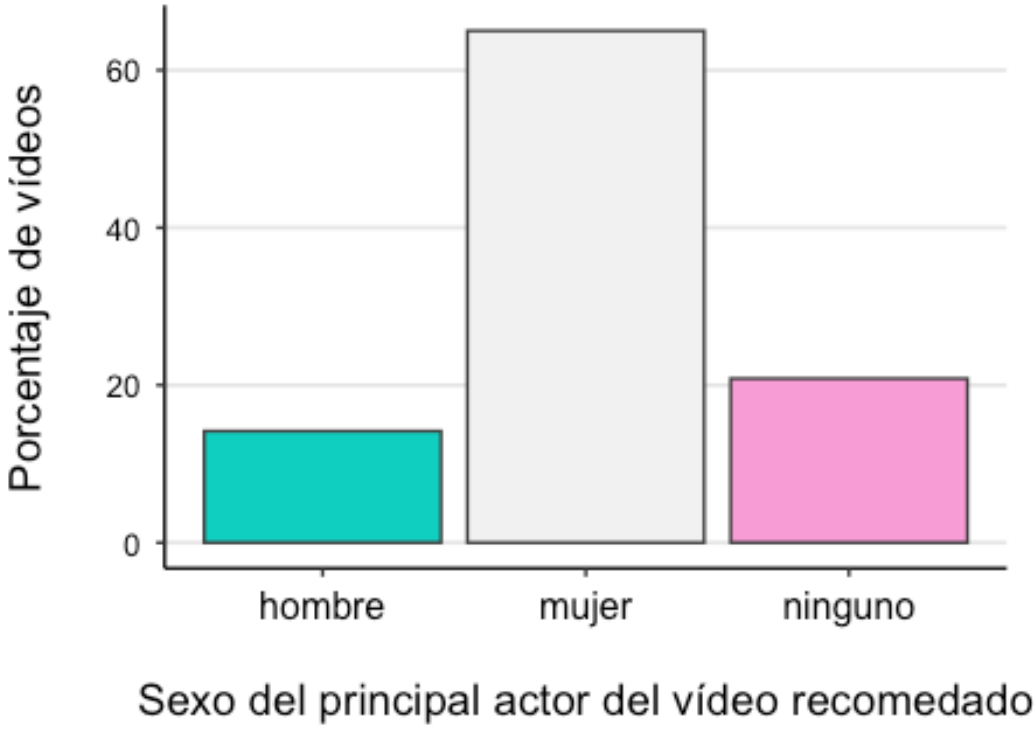
Temáticas de los vídeos recomendados

```
ggplot(data = tiktok) +
  geom_bar(aes(x=fct_infreq(tematica), y=after_stat(count/sum(count)*100)),
           fill=hcl.colors(7, palette="Cyan-Mage"),
           col="grey25") +
  labs(title=" ",
       x="\nTemática del vídeo recomedado",
       y="Porcentaje de vídeos\n")
```



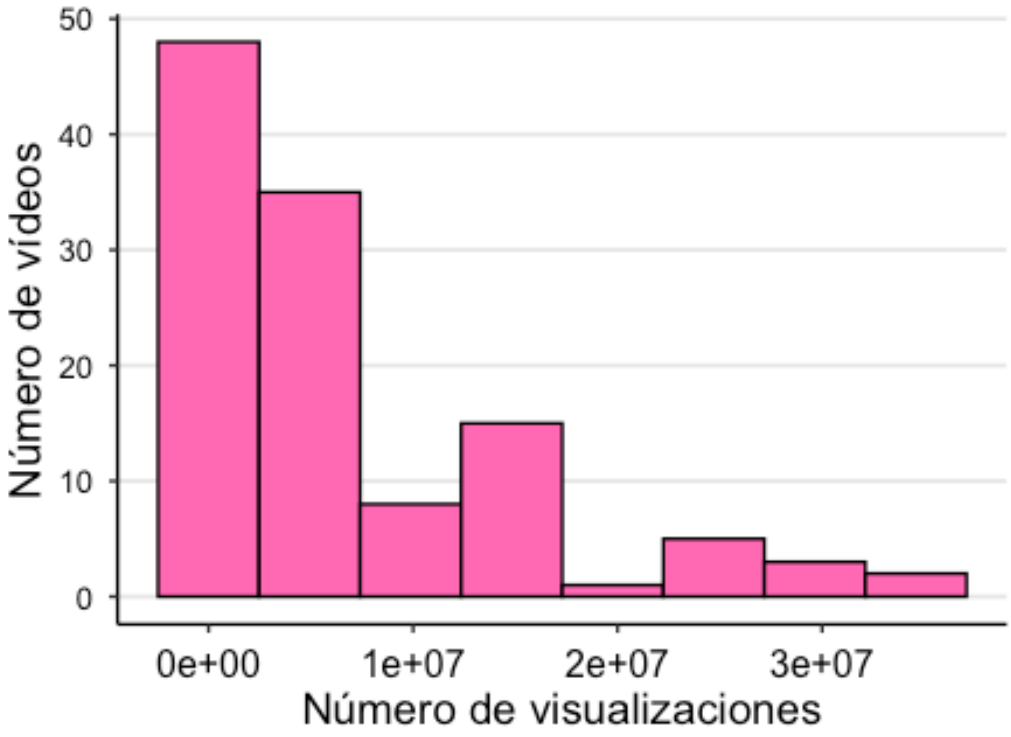
Sexo del actor principal que aparece en el vídeo recomendado

```
ggplot(data = tiktok) +
  geom_bar(aes(x=sexvid, y=after_stat(count/sum(count)*100)),
           fill=hcl.colors(3, palette="Cyan-Mage"),
           col="grey25") +
  labs(title=" ",
       x="\nSexo del principal actor del vídeo recomedado",
       y="Porcentaje de vídeos\n")
```



Visualizaciones

```
ggplot(data=tiktok, aes(x=viewsvid)) +
  geom_histogram(bins = 8, fill="hotpink", color="black") +
  labs(title=" ",
    x="Número de visualizaciones",
    y="Número de vídeos")
```



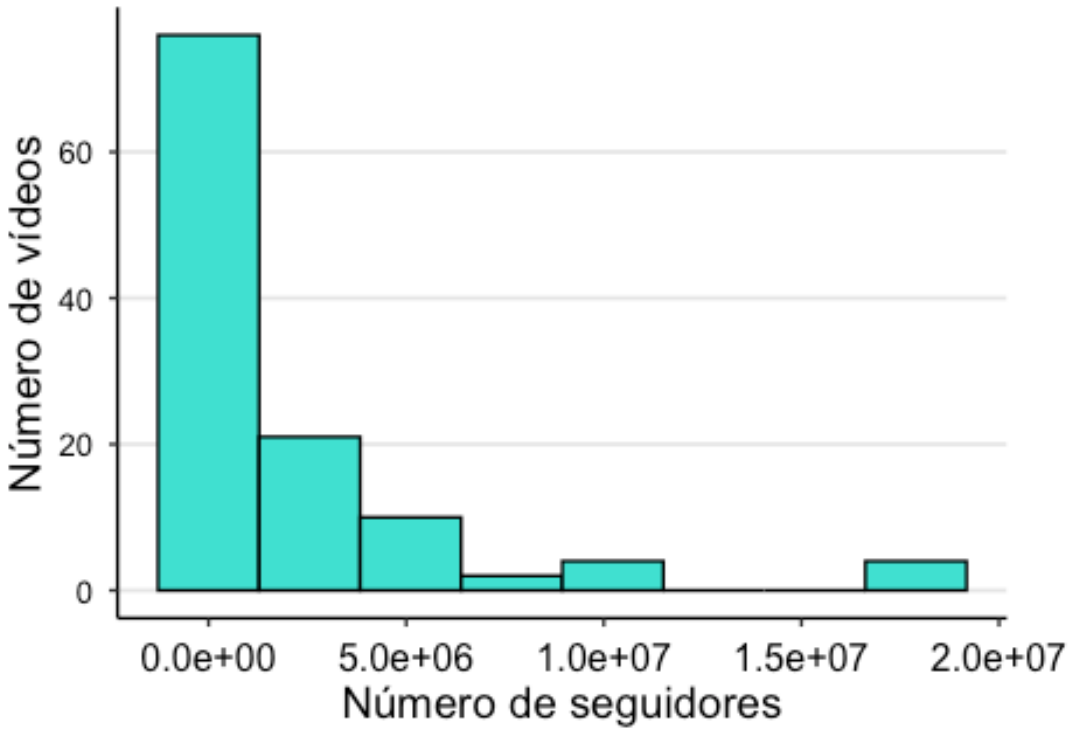
```
summary(tiktok$viewsvid)
```

```
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##  208400  1600000  3200000  7034591  9300000 34800000        3
```

3.1.2. Características y métricas de las cuentas que han generado los vídeos recomendados

Seguidores

```
ggplot(data=tiktok, aes(x=seguidores)) +
  geom_histogram(bins = 8, fill="turquoise", color="black") +
  labs(title=" ",
    x="Número de seguidores",
    y="Número de vídeos")
```




```
summary(tiktok$seguidores)
```

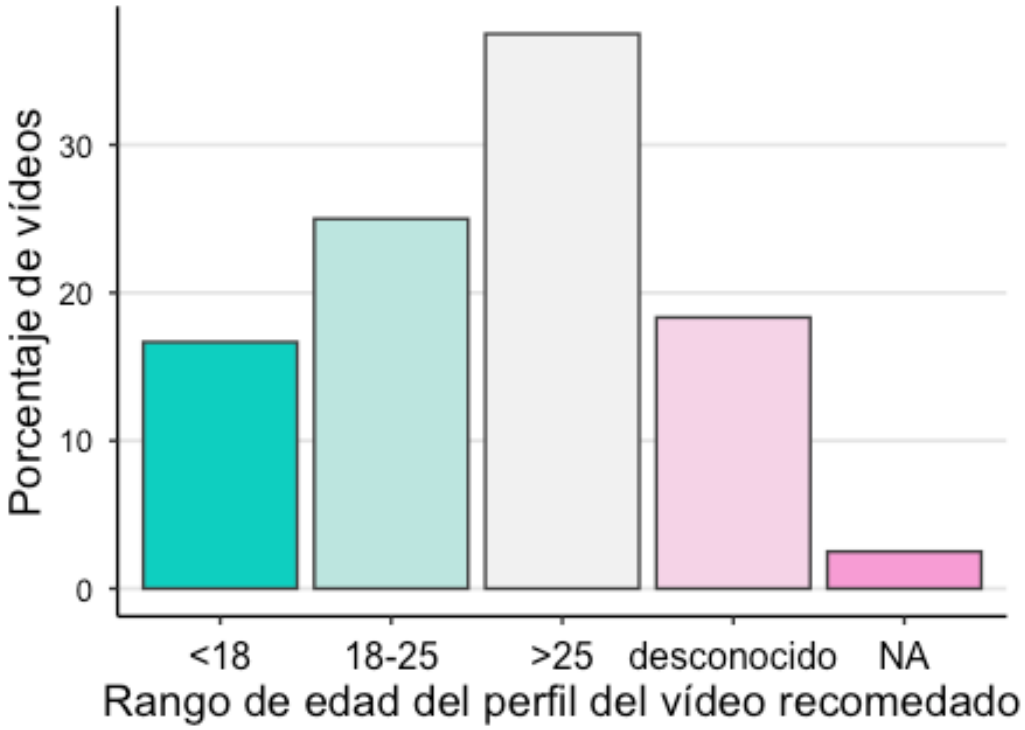
```
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     670   105300   519400  2029888  1600000 17900000        3
```

Rango de edad

```
levels(tiktok$rango)
```

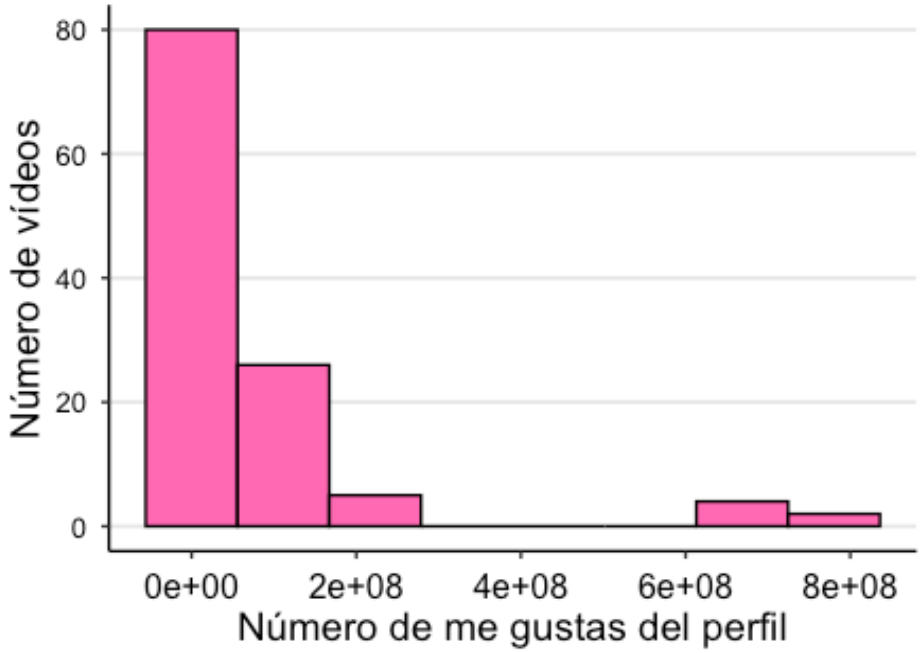
```
## [1] "<18"         "18-25"       ">25"         "desconocido"
```

```
tiktok$rango = factor(tiktok$rango, levels=c("<18", "18-25", ">25", "desconocido", "NA"
))
ggplot(data = tiktok) +
  geom_bar(aes(x=rango, y=after_stat(count/sum(count)*100)),
           fill=hcl.colors(5, palette="Cyan-Mage"),
           col="grey25") +
  labs(title=" ",
       x="Rango de edad del perfil del vídeo recomedado",
       y="Porcentaje de vídeos")
```



Número de me gustas del perfil

```
ggplot(data=tiktok, aes(x=nmegustas)) +
  geom_histogram(bins = 8, fill="hotpink", color="black") +
  labs(title=" ",
    x="Número de me gustas del perfil",
    y="Número de vídeos")
```




```
summary(tiktok$nmegustas)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##     89500   1600000  15800000  74959062  94500000 780400000         3
```

Check azul

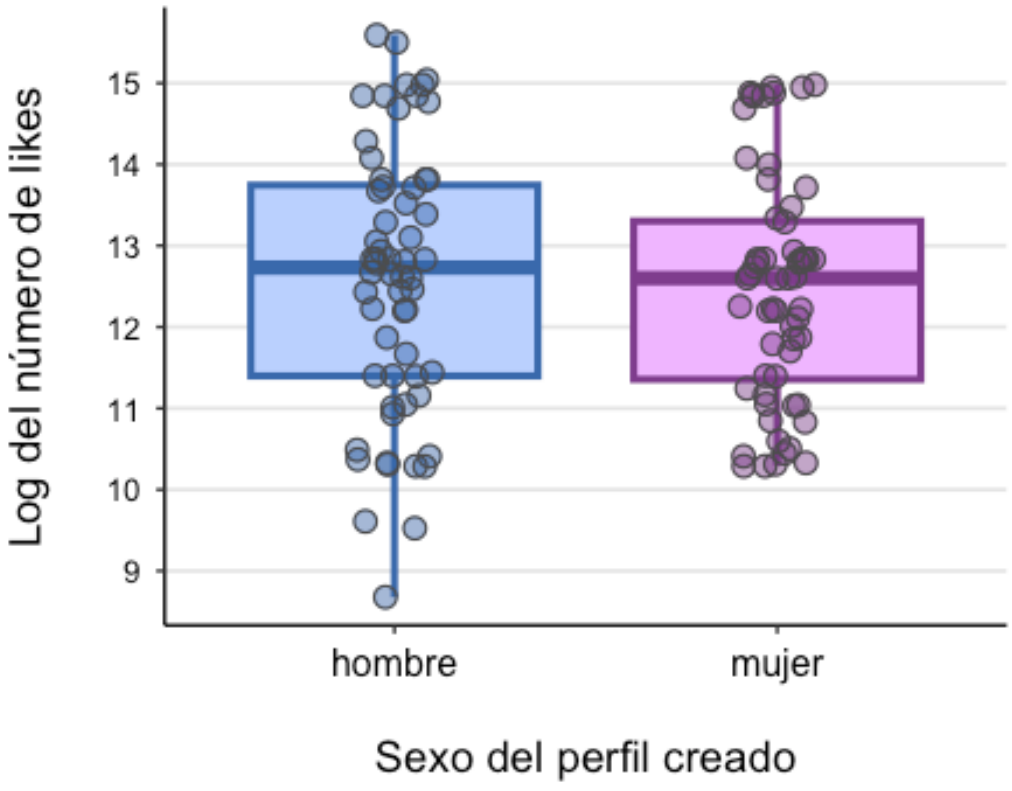
```
ggplot(data = tiktok) +
  geom_bar(aes(x=checkazul, y=after_stat(count/sum(count)*100)),
           fill=hcl.colors(2, palette="Cyan-Mage"),
           col="grey25") +
  labs(
      x="Perfiles con check azul",
      y="Porcentaje de vídeos")
```

3.2. Estudio de relaciones entre variables

3.2.1 Entre métricas de los perfiles creados y otras métricas

Sexo del perfil creado y me gustas del vídeo

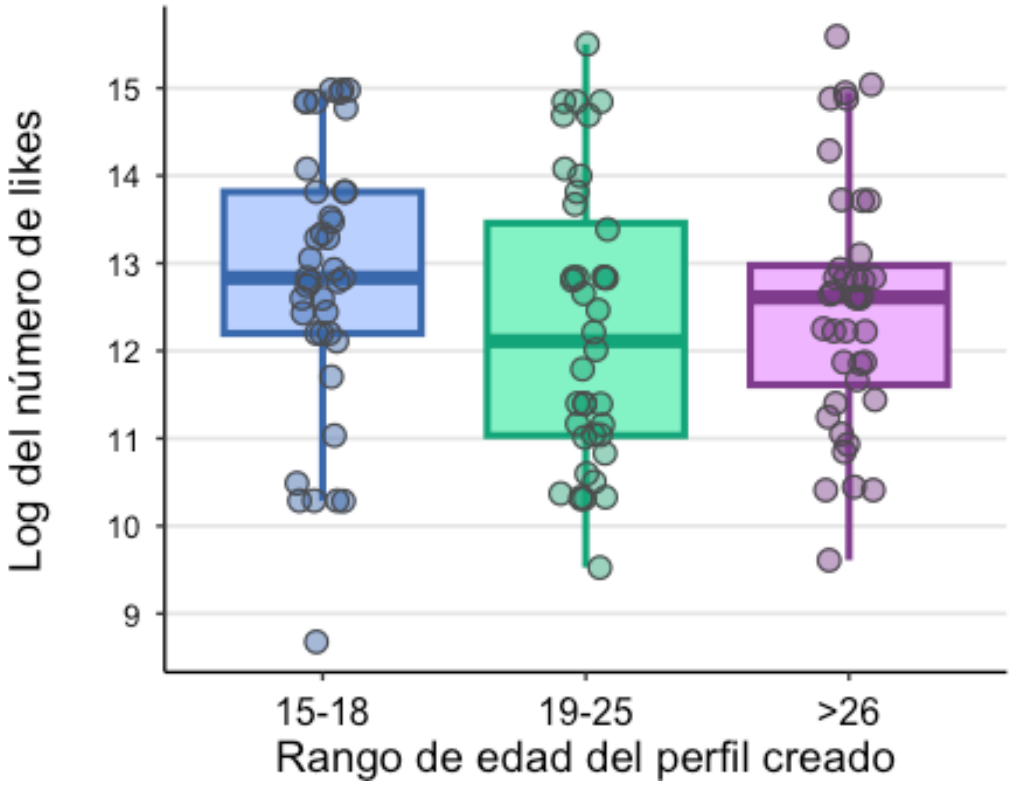
```
## custom colors
my_pal2 <- rcartocolor::carto_pal(n = 12, name = "Bold")[c(3,1)]

g1 <- ggplot(data=tiktok, aes(x = sexopc, y = log(mgvid), color = sexopc, fill = sexopc)
) +
  scale_y_continuous(breaks=5:16) +
  scale_color_manual(values = my_pal2, guide = "none") +
  scale_fill_manual(values = my_pal2, guide = "none") +
  labs(
      x="\nSexo del perfil creado",
      y="Log del número de likes\n")

g1 +
  geom_boxplot(
    aes(fill = sexopc, fill = after_scale(colorspace::lighten(fill, .7))),
    size = 1.1, outlier.shape = NA
  ) +
  geom_point(
    position = position_jitter(width = .1, seed = 0),
    size = 3, alpha = .5
  ) +
  geom_point(
    position = position_jitter(width = .1, seed = 0),
    size = 3, stroke = .6, shape = 1, color = "gray30"
  )
```



```
tapply(tiktok$mgvid , tiktok$sexopc, summary)

## $hombre
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5868   89150  338500  873081  933550 5900000
##
## $mujer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29400   85625  296850  678348  599225 3200000

fit8 <- aov(lm(log(tiktok$mgvid)~tiktok$sexopc))
summary(fit8)

##               Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$sexopc   1    0.3   0.301   0.127  0.722
## Residuals     118  279.4   2.368
```

Rango de edad del perfil creado y me gustas del vídeo

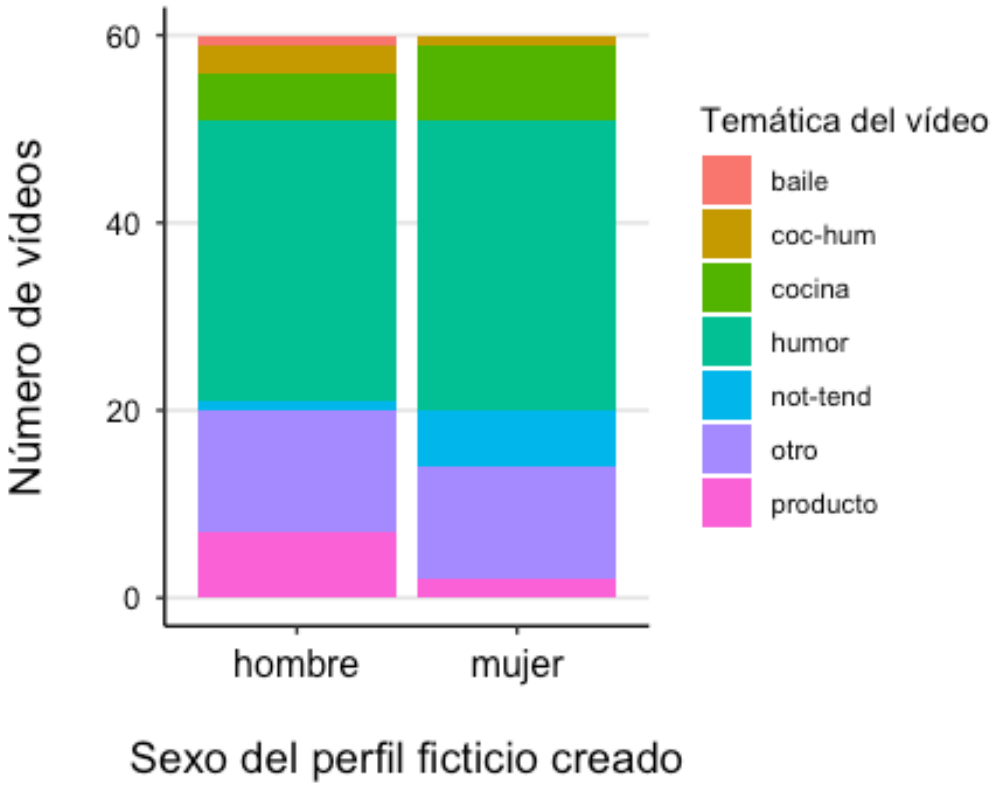
```
## custom colors
my_pal2 <- rcartocolor::carto_pal(n = 12, name = "Bold")[c(3,2,1)]

tiktok$rangopc = factor(tiktok$rangopc, levels=c("15-18", "19-25", ">26"))

g1 <- ggplot(data=tiktok, aes(x = rangopc, y = log(mgvid), color = rangopc, fill = rango
pc)) +
  scale_y_continuous(breaks=5:16) +
  scale_color_manual(values = my_pal2, guide = "none") +
  scale_fill_manual(values = my_pal2, guide = "none") +
  labs(x="Rango de edad del perfil creado",
       y="Log del número de likes\n")

g1 +
  geom_boxplot(
    aes(fill = rangopc, fill = after_scale(colorspace::lighten(fill, .7))),
    size = 1.1, outlier.shape = NA
  ) +
  geom_point(
    position = position_jitter(width = .1, seed = 0),
    size = 3, alpha = .5
  ) +
  geom_point(
    position = position_jitter(width = .1, seed = 0),
    size = 3, stroke = .6, shape = 1, color = "gray30"
  )
```




```
tapply(tiktok$mgvid , tiktok$rangopc, summary)

## $`15-18`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5868  198600  374550  904094 1000000 3200000
##
## $`19-25`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13700   61875  183050  697692  705400 5400000
##
## $`>26`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14900  110875  299500  725358  432350 5900000

fit9 <- aov(lm(log(tiktok$mgvid)~tiktok$rangopc))
summary(fit9)

##                Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$rangopc   2   6.54   3.272   1.402   0.25
## Residuals      117 273.16   2.335
```

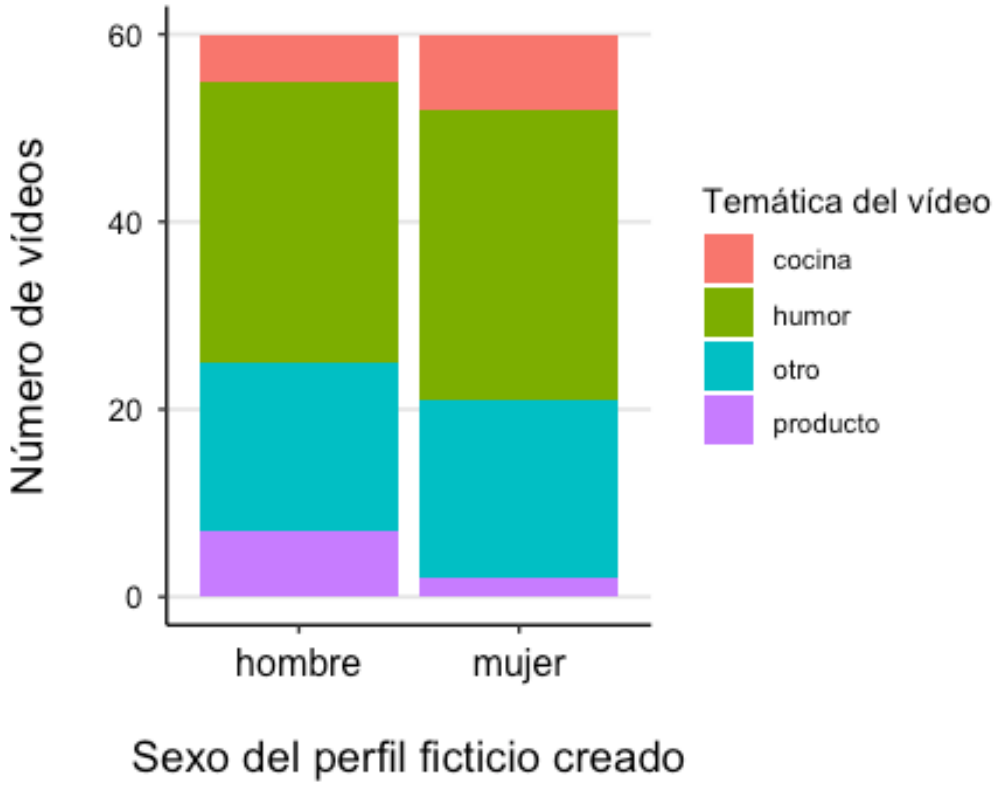
Sexo del perfil creado y la temática del vídeo recomendado

```
ggplot(data = tiktok) +
  geom_bar(aes(x=sexopc, fill=tematica)) +
  labs(
      x="\nSexo del perfil ficticio creado",
      y="Número de vídeos\n",
      fill="Temática del vídeo")
```




```
t2 = table(tiktok$sexopc, tiktok$tematica)
t2

##
##          baile coc-hum cocina humor not-tend otro producto
##   hombre     1       3      5    30        1   13        7
##   mujer      0       1      8    31        6   12        2

round(prop.table(t2, margin=1)*100, 2)

##
##          baile coc-hum cocina humor not-tend  otro producto
##   hombre  1.67    5.00   8.33 50.00     1.67 21.67    11.67
##   mujer   0.00    1.67  13.33 51.67    10.00 20.00     3.33

chisq.test(tiktok$sexopc, tiktok$tematica)

## Warning in chisq.test(tiktok$sexopc, tiktok$tematica): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  tiktok$sexopc and tiktok$tematica
## X-squared = 9.0979, df = 6, p-value = 0.1681

chisq.test(tiktok$sexopc, tiktok$tematica)$expected
```

```
## Warning in chisq.test(tiktok$sexopc, tiktok$tematica): Chi-squared
## approximation may be incorrect

##              tiktok$tematica
## tiktok$sexopc baile coc-hum cocina humor not-tend otro producto
##       hombre   0.5       2    6.5  30.5      3.5 12.5      4.5
##       mujer    0.5       2    6.5  30.5      3.5 12.5      4.5

tema <- tiktok$tematica
tema[tema=="baile"]<-"otro"
tema[tema=="coc-hum"]<-"otro"
tema[tema=="not-tend"]<-"otro"
tema <- as.factor(tema)
tiktok$tema <- droplevels(tema)
table(tiktok$tema)

##
##   cocina    humor     otro producto
##       13       61       37        9

ggplot(data = tiktok) +
  geom_bar(aes(x=sexopc, fill=tema)) +
  labs(x="\nSexo del perfil ficticio creado",
       y="Número de vídeos\n",
       fill="Temática del vídeo")
```



```r
t3 = table(tiktok$sexopc, tiktok$tema)
t3
```

```
##
##          cocina humor otro producto
##   hombre      5    30   18        7
##   mujer       8    31   19        2
```

```r
round(prop.table(t3, margin=1)*100, 2)
```

```
##
##          cocina humor  otro producto
##   hombre   8.33 50.00 30.00    11.67
##   mujer   13.33 51.67 31.67     3.33
```

```r
chisq.test(tiktok$sexopc, tiktok$tema)
```

```
## Warning in chisq.test(tiktok$sexopc, tiktok$tema): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tiktok$sexopc and tiktok$tema
## X-squared = 3.5135, df = 3, p-value = 0.319
```

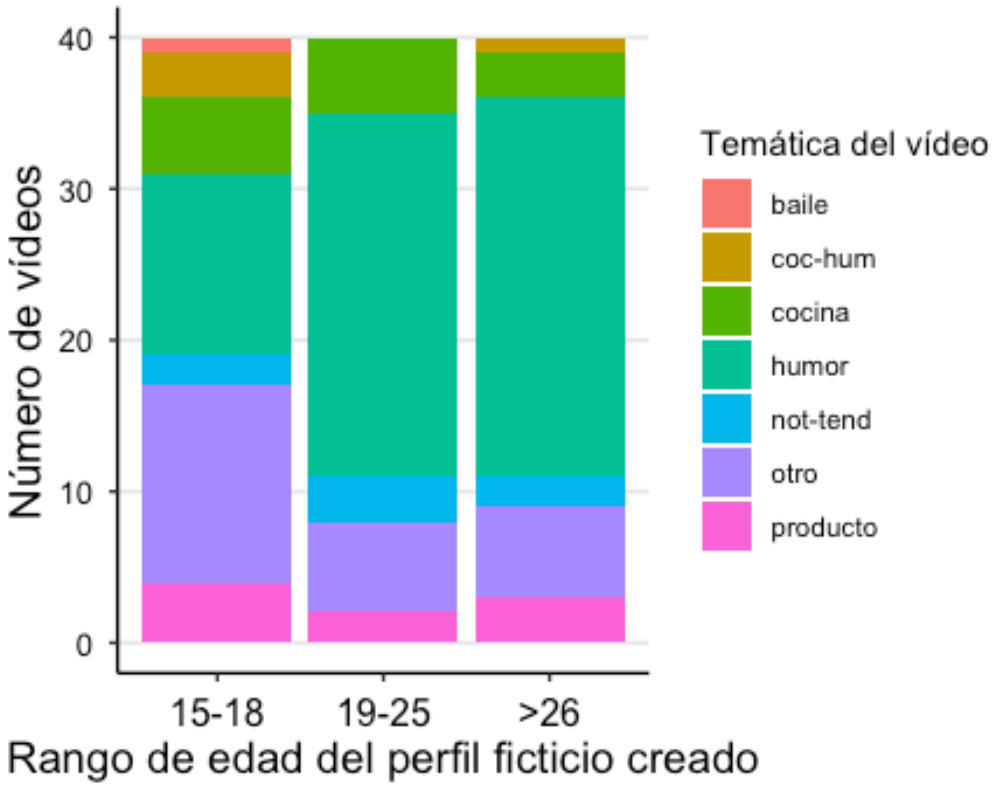
```r
chisq.test(tiktok$sexopc, tiktok$tema)$expected
```

```
## Warning in chisq.test(tiktok$sexopc, tiktok$tema): Chi-squared approximation
## may be incorrect
```

```
##              tiktok$tema
## tiktok$sexopc cocina humor otro producto
##       hombre    6.5  30.5 18.5      4.5
##       mujer     6.5  30.5 18.5      4.5
```

Rango de edad del perfil creado y la temática del vídeo recomendado

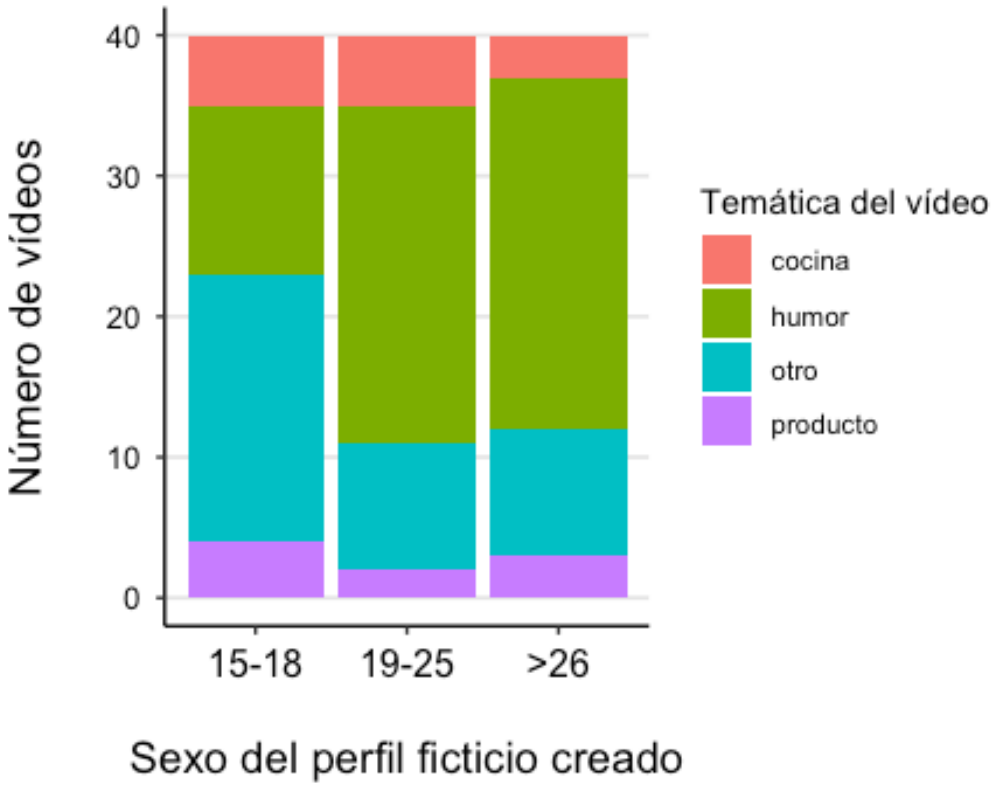
```r
ggplot(data = tiktok) +
  geom_bar(aes(x=rangopc, fill=tematica)) +
  labs(x="Rango de edad del perfil ficticio creado",
       y="Número de vídeos",
       fill="Temática del vídeo")
```




```
t4 = table(tiktok$rangopc, tiktok$tematica)
t4

##
##        baile coc-hum cocina humor not-tend otro producto
##   15-18     1      3     5    12        2   13        4
##   19-25     0      0     5    24        3    6        2
##   >26       0      1     3    25        2    6        3

round(prop.table(t4, margin=1)*100, 2)

##
##        baile coc-hum cocina humor not-tend otro producto
##   15-18   2.5    7.5   12.5  30.0      5.0 32.5     10.0
##   19-25   0.0    0.0   12.5  60.0      7.5 15.0      5.0
##   >26     0.0    2.5    7.5  62.5      5.0 15.0      7.5

chisq.test(tiktok$rangopc, tiktok$tematica)

## Warning in chisq.test(tiktok$rangopc, tiktok$tematica): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  tiktok$rangopc and tiktok$tematica
## X-squared = 16.135, df = 12, p-value = 0.1851

chisq.test(tiktok$rangopc, tiktok$tematica)$expected
```

```
## Warning in chisq.test(tiktok$rangopc, tiktok$tematica): Chi-squared
## approximation may be incorrect

##              tiktok$tematica
## tiktok$rangopc     baile  coc-hum    cocina    humor not-tend     otro producto
##          15-18 0.3333333 1.333333 4.333333 20.33333 2.333333 8.333333        3
##          19-25 0.3333333 1.333333 4.333333 20.33333 2.333333 8.333333        3
##          >26   0.3333333 1.333333 4.333333 20.33333 2.333333 8.333333        3
```

```
ggplot(data = tiktok) +
  geom_bar(aes(x=rangopc, fill=tema)) +
  labs(x="\nSexo del perfil ficticio creado",
       y="Número de vídeos\n",
       fill="Temática del vídeo")
```




```
t5 = table(tiktok$rangopc, tiktok$tema)
t5

##
##         cocina humor otro producto
##   15-18      5    12   19        4
##   19-25      5    24    9        2
##   >26        3    25    9        3

round(prop.table(t5, margin=1)*100, 2)

##
##         cocina humor otro producto
##   15-18   12.5  30.0 47.5     10.0
```

```
##   19-25   12.5  60.0 22.5      5.0
##   >26      7.5  62.5 22.5      7.5
```

```
chisq.test(tiktok$rangopc, tiktok$tema)
```

```
## Warning in chisq.test(tiktok$rangopc, tiktok$tema): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tiktok$rangopc and tiktok$tema
## X-squared = 11.835, df = 6, p-value = 0.06575
```

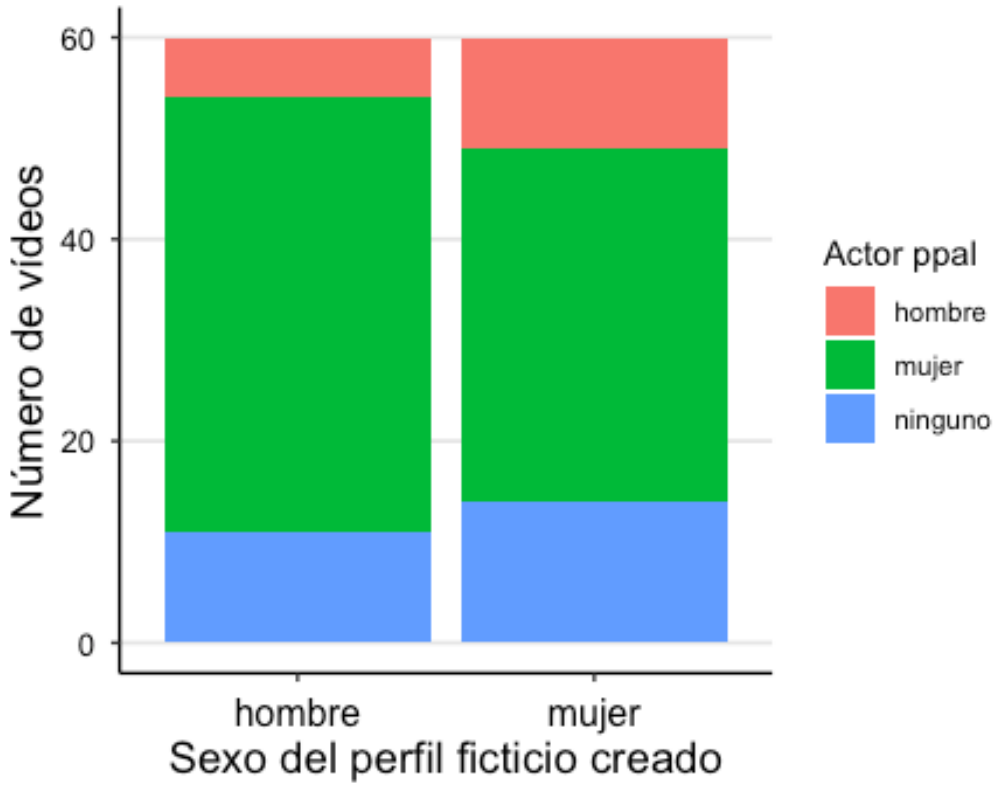
```
chisq.test(tiktok$rangopc, tiktok$tema)$expected
```

```
## Warning in chisq.test(tiktok$rangopc, tiktok$tema): Chi-squared approximation
## may be incorrect
```

```
##               tiktok$tema
## tiktok$rangopc   cocina    humor     otro producto
##          15-18 4.333333 20.33333 12.33333        3
##          19-25 4.333333 20.33333 12.33333        3
##          >26   4.333333 20.33333 12.33333        3
```

Sexo del perfil creado y del actor principal

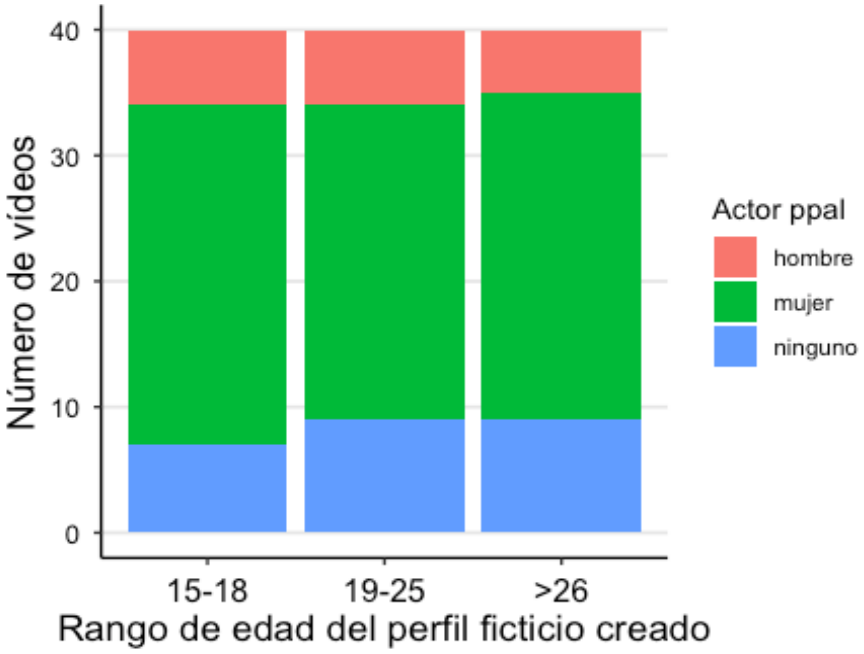
```
ggplot(data = tiktok) +
  geom_bar(aes(x=sexopc, fill=sexvid)) +
  labs(x="Sexo del perfil ficticio creado",
       y="Número de vídeos",
       fill="Actor ppal")
```



```
t2 = table(tiktok$sexopc, tiktok$sexvid)
t2

##
##          hombre mujer ninguno
##   hombre      6    43      11
##   mujer      11    35      14

round(prop.table(t2, margin=1)*100, 2)

##
##          hombre mujer ninguno
##   hombre  10.00 71.67   18.33
##   mujer   18.33 58.33   23.33

chisq.test(tiktok$sexopc, tiktok$sexvid)

##
##  Pearson's Chi-squared test
##
## data:  tiktok$sexopc and tiktok$sexvid
## X-squared = 2.6511, df = 2, p-value = 0.2657

chisq.test(tiktok$sexopc, tiktok$sexvid)$expected

##              tiktok$sexvid
## tiktok$sexopc hombre mujer ninguno
##       hombre    8.5    39    12.5
##       mujer     8.5    39    12.5
```

Rango de edad del perfil creado y sexo del actor principal

```
ggplot(data = tiktok) +
  geom_bar(aes(x=rangopc, fill=sexvid)) +
  labs(x="Rango de edad del perfil ficticio creado",
       y="Número de vídeos",
       fill="Actor ppal")
```

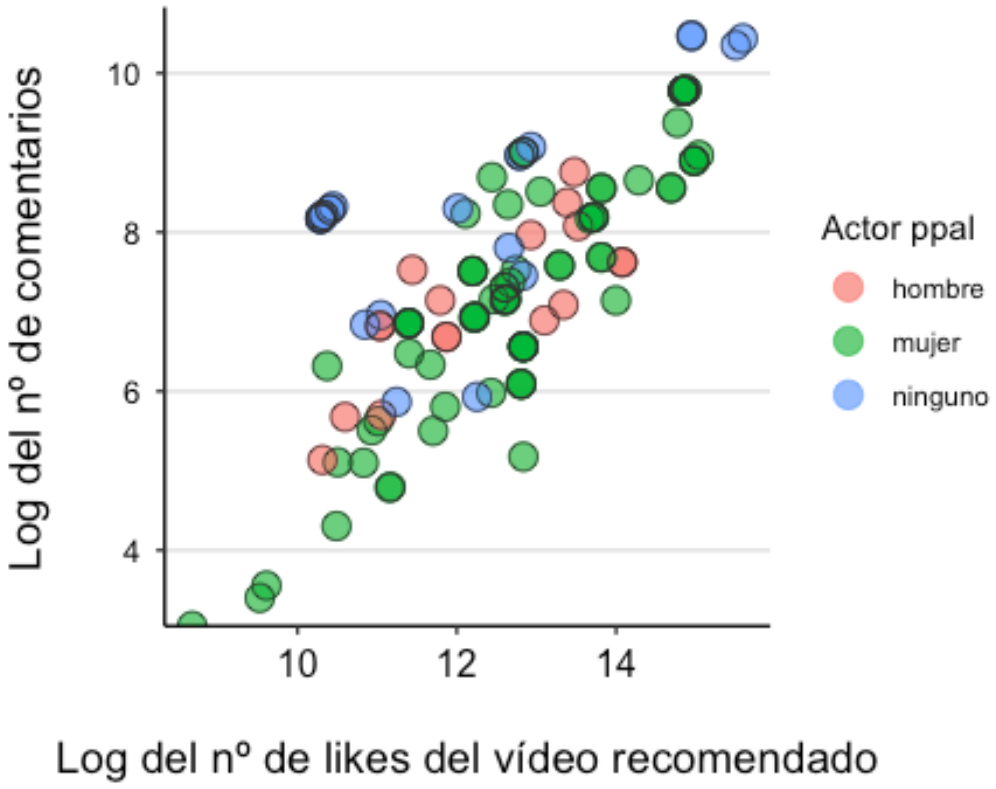


```
t4 = table(tiktok$rangopc, tiktok$sexvid)
t4

##
##        hombre mujer ninguno
##   15-18      6    27       7
##   19-25      6    25       9
##   >26        5    26       9

round(prop.table(t4, margin=1)*100, 3)

##
##        hombre mujer ninguno
##   15-18   15.0  67.5    17.5
##   19-25   15.0  62.5    22.5
##   >26     12.5  65.0    22.5

chisq.test(tiktok$rangopc, tiktok$sexvid)

##
##  Pearson's Chi-squared test
##
## data:  tiktok$rangopc and tiktok$sexvid
## X-squared = 0.51457, df = 4, p-value = 0.9721

chisq.test(tiktok$rangopc, tiktok$sexvid)$expected

##              tiktok$sexvid
## tiktok$rangopc   hombre mujer  ninguno
##         15-18 5.666667    26 8.333333
##         19-25 5.666667    26 8.333333
##         >26   5.666667    26 8.333333
```

3.2.2. Entre métricas del vídeo recomendado

Comentarios y me gustas del vídeo

```
g <- ggplot(data=tiktok, aes(x = log(mgvid), y = log(comentarios), color = sexvid)) +
  scale_y_continuous(breaks=seq(0, 12, by=2)) +
  labs(x="\nLog del nº de likes del vídeo recomendado",
       y="Log del nº de comentarios\n",
       color = "Actor ppal")

g +
  geom_point(size = 4, alpha = .7) +
  geom_point(size = 4, stroke = .4, shape = 1, color = "gray20")
```

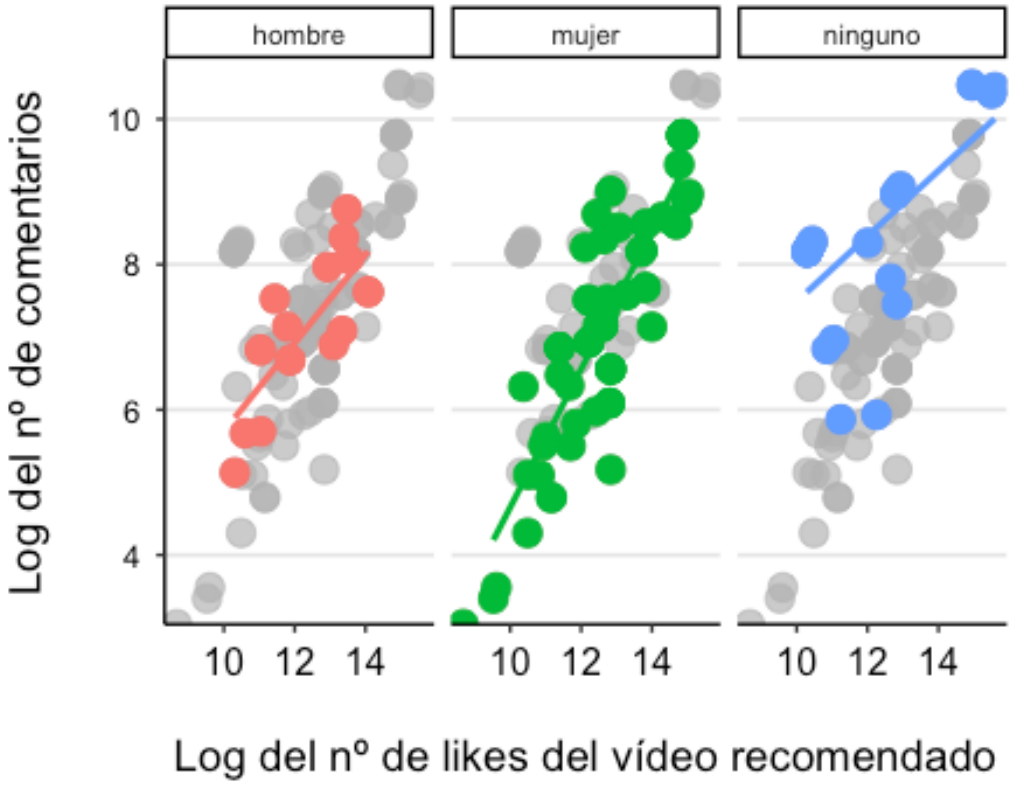



```
cor(log(tiktok$mgvid+1), log(tiktok$comentarios+1))
```

```
## [1] 0.6922307
```

```
g <- ggplot(data=tiktok, aes(x = log(mgvid), y = log(comentarios))) +
  scale_y_continuous(breaks=seq(0, 12, by=2)) +
  labs(x="\nLog del nº de likes del vídeo recomendado",
       y="Log del nº de comentarios\n")

# Base de datos auxiliar para dibujar los puntos en gris (sin diferenciar por el sexo del actor principal del video)
aux <- dplyr::select(tiktok, -sexvid)

g +
  geom_point(data = aux, size = 4, alpha = .7, color = "gray70") +
  geom_point(aes(colour=sexvid), size = 4)+
  geom_smooth(aes(colour=sexvid), method="lm", se=FALSE)+
  facet_wrap(~sexvid)+
  theme(legend.position="none") #"bottom"
```

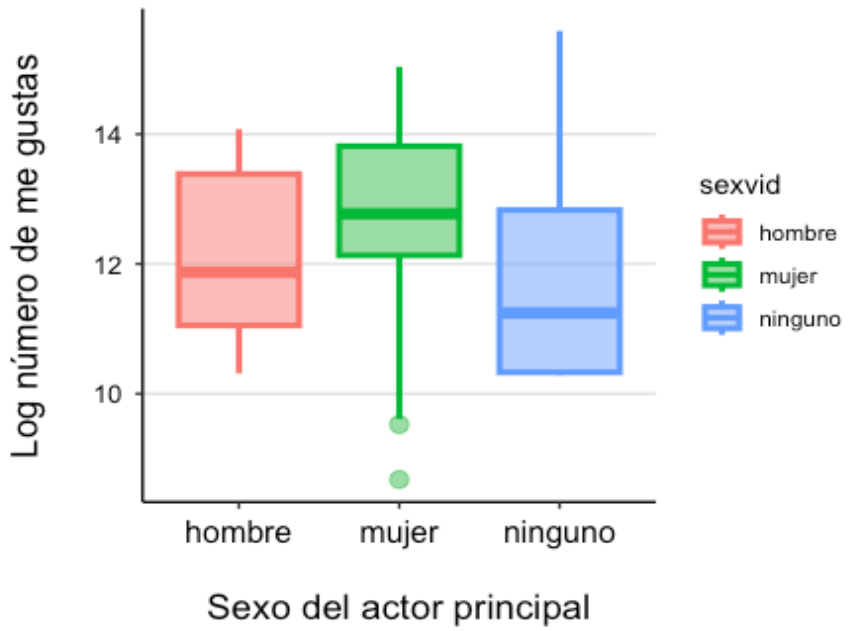
```
## `geom_smooth()` using formula = 'y ~ x'
```



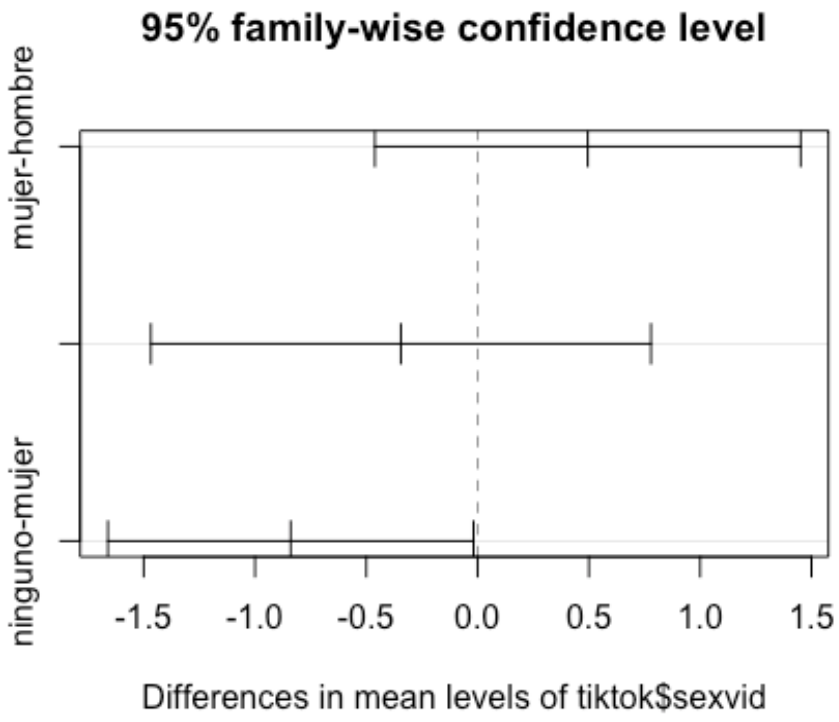

```
fit <- lm(log(mgvid+1)~log(comentarios+1)*sexvid, data=tiktok, na.action = na.exclude
)
summary(fit)

##
## Call:
## lm(formula = log(mgvid + 1) ~ log(comentarios + 1) * sexvid,
##     data = tiktok, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7215 -0.5691  0.1341  0.5628  2.6109
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.04133    1.65994   3.037  0.00296 **
## log(comentarios + 1)              1.02159    0.23193   4.405  2.4e-05 ***
## sexvidmujer                       2.40487    1.71855   1.399  0.16442
## sexvidninguno                    -1.01198    2.08471  -0.485  0.62830
## log(comentarios + 1):sexvidmujer  -0.28639    0.23947  -1.196  0.23421
## log(comentarios + 1):sexvidninguno -0.07529    0.27579  -0.273  0.78534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9025 on 114 degrees of freedom
## Multiple R-squared:  0.6681, Adjusted R-squared:  0.6535
## F-statistic: 45.89 on 5 and 114 DF,  p-value: < 2.2e-16
```

Actor principal y me gustas del vídeo

```
g2 <- ggplot(data=tiktok, aes(x = sexvid, y = log(mgvid), color = sexvid, fill = sexvid)) +
  scale_y_continuous() +
  labs(x="\nSexo del actor principal",
       y="Log número de me gustas\n")

g2 +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```

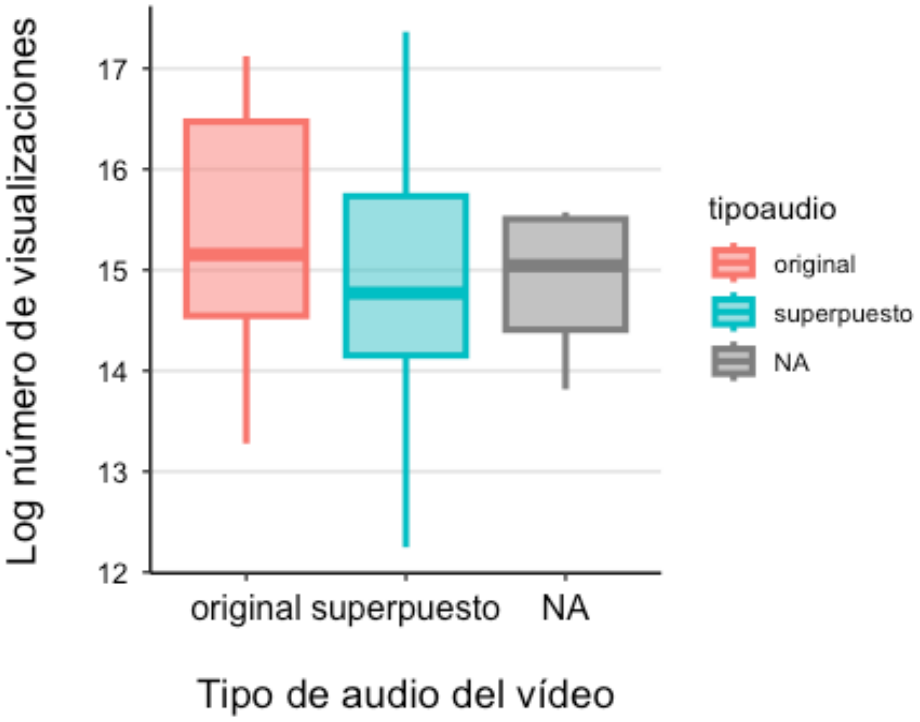



```
tapply(tiktok$mgvid , tiktok$sexvid, summary)

## $hombre
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30100   63100  143000  411729  651400 1300000
##
## $mujer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5868  185850  352650  839950 1000000 3400000
##
## $ninguno
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29300   30600   76400  822812  373400 5900000

fit <- aov(lm(log(tiktok$mgvid)~tiktok$sexvid))
summary(fit)

##               Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$sexvid   2  14.59   7.293   3.219 0.0436 *
## Residuals     117 265.12   2.266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

intervalos=TukeyHSD(fit)
plot(intervalos)
```



Tipo de audio y visualizaciones

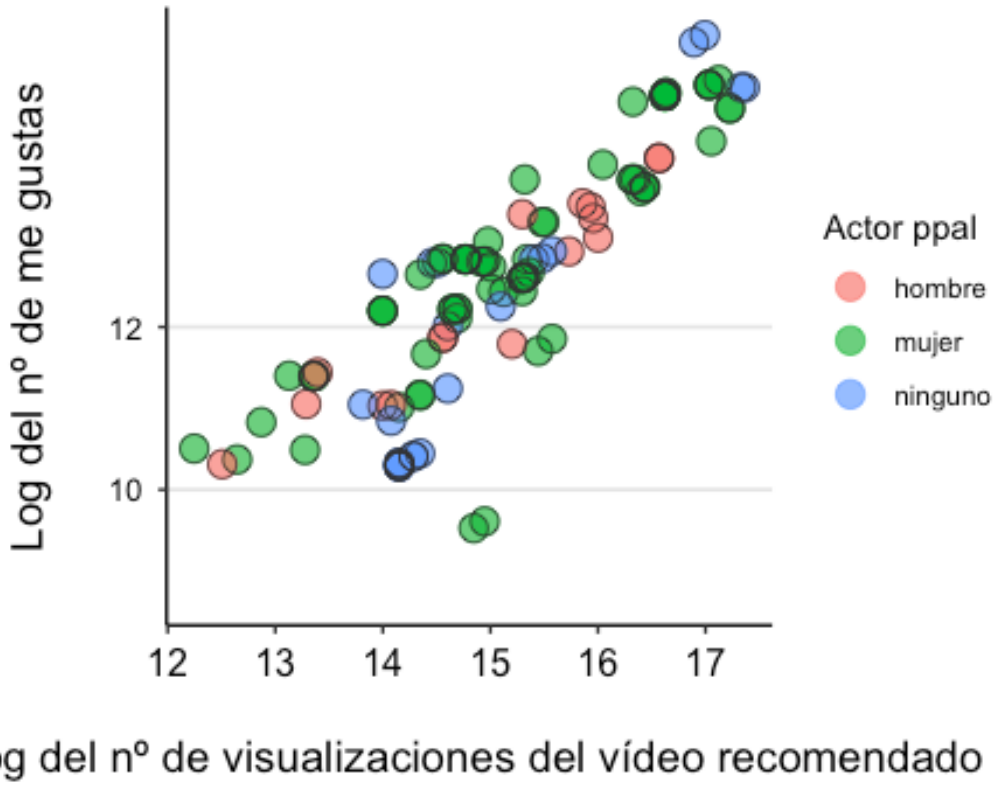
```
g <- ggplot(data=tiktok, aes(x = tipoaudio, y = log(viewsvid), color = tipoaudio, fill = tip
oaudio)) +
  scale_y_continuous() +
  labs(x="\nTipo de audio del vídeo",
       y="Log número de visualizaciones\n")

g +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```



```
tapply(tiktok$viewsvid , tiktok$tipoaudio, summary)

## $original
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   583700  2075000  3850000  8568992 14275000 27300000        3
##
## $superpuesto
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   208400  1400000  2600000  6496929  6800000 34800000

fit2 <- aov(lm(log(tiktok$viewsvid)~tiktok$tipoaudio))
summary(fit2)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$tipoaudio   1   5.12   5.121   3.647 0.0588 .
## Residuals        111 155.85   1.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 7 observations deleted due to missingness
```

Visualizaciones y me gustas del vídeo

```
g <- ggplot(data=tiktok, aes(x = log(viewsvid), y = log(mgvid), color = sexvid)) +
  scale_y_continuous(breaks=seq(0, 12, by=2)) +
  labs(x="\nLog del nº de visualizaciones del vídeo recomendado",
       y="Log del nº de me gustas\n",
       color = "Actor ppal")

g +
  geom_point(size = 4, alpha = .7) +
  geom_point(size = 4, stroke = .4, shape = 1, color = "gray20")
```

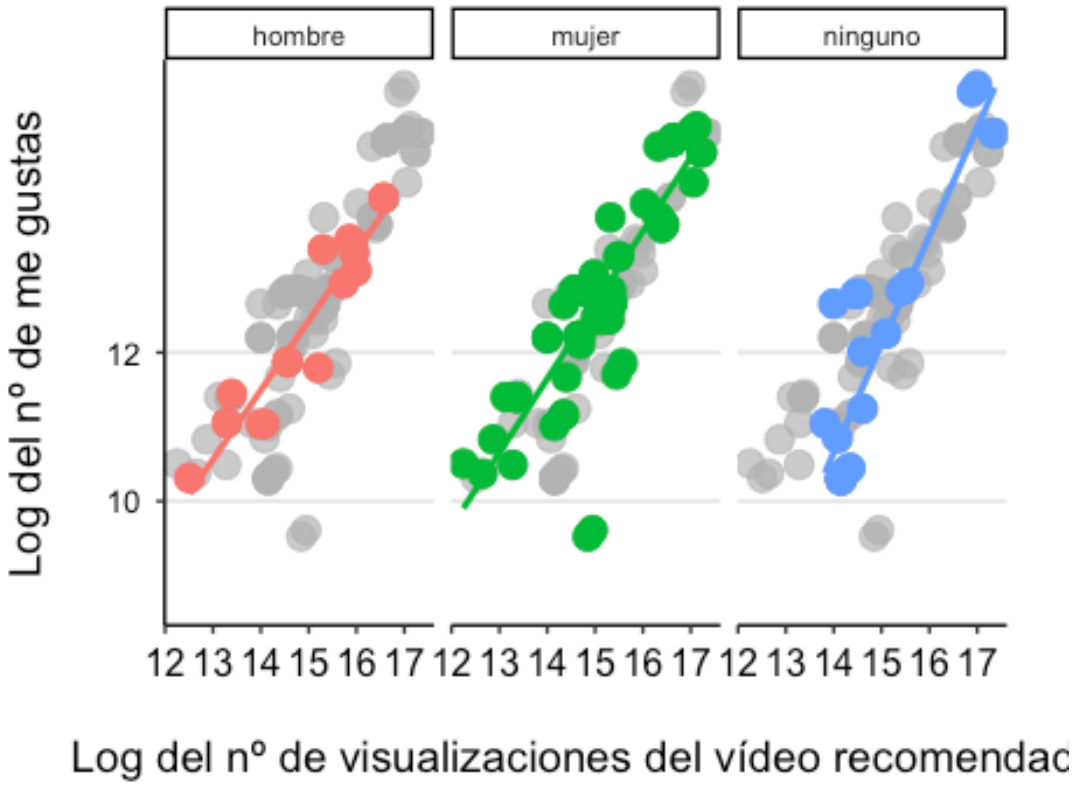


```r
cor(log(tiktok$mgvid+1), log(tiktok$viewsvid+1), use="complete.obs")
```

```
## [1] 0.8693661
```

```r
g <- ggplot(data=tiktok, aes(x = log(viewsvid), y = log(mgvid))) +
  scale_y_continuous(breaks=seq(0, 12, by=2)) +
  labs(x="\nLog del nº de visualizaciones del vídeo recomendado",
       y="Log del nº de me gustas\n")

# Base de datos auxiliar para dibujar los puntos en gris (sin diferenciar por el sexo del actor principal del video)
aux <- dplyr::select(tiktok, -sexvid)

g +
  geom_point(data = aux, size = 4, alpha = .7, color = "gray70") +
  geom_point(aes(colour=sexvid), size = 4)+
  geom_smooth(aes(colour=sexvid), method="lm", se=FALSE)+
  facet_wrap(~sexvid)+
  theme(legend.position="none") #"bottom"
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



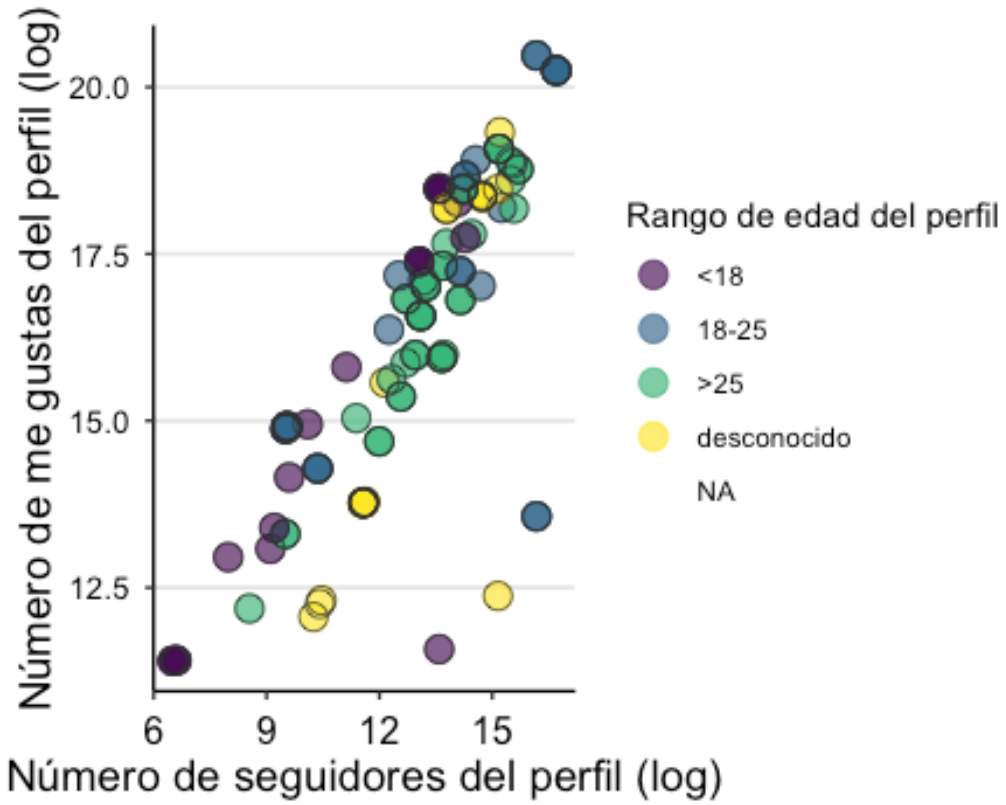
```r
fit3 <- lm(log(viewsvid+1)~log(mgvid+1)*sexvid, data=tiktok, na.action = na.exclude)
summary(fit3)
```

```
##
## Call:
## lm(formula = log(viewsvid + 1) ~ log(mgvid + 1) * sexvid, data = tiktok,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28721 -0.34361  0.03396  0.26903  2.17985
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.0646     1.4706   2.084  0.03947 *
## log(mgvid + 1)                 0.9604     0.1181   8.131 6.66e-13 ***
## sexvidmujer                    2.4392     1.5910   1.533  0.12809
## sexvidninguno                  5.1214     1.6599   3.085  0.00257 **
## log(mgvid + 1):sexvidmujer    -0.2045     0.1271  -1.609  0.11042
## log(mgvid + 1):sexvidninguno  -0.3993     0.1342  -2.975  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5592 on 111 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.7874, Adjusted R-squared:  0.7779
## F-statistic: 82.24 on 5 and 111 DF,  p-value: < 2.2e-16
```

3.2.3 Entre las métricas de las cuentas de los vídeos

Seguidores y número de me gustas del perfil

```r
g <- ggplot(data=tiktok, aes(x = log(seguidores), y = log(nmegustas), color = rango))
+
  labs(x="Número de seguidores del perfil (log)",
       y="Número de me gustas del perfil (log)",
       color = "Rango de edad del perfil")

# g+geom_point(size = 3, alpha = .7,  stroke = .4, shape = 1, color = "gray20")

g +
  geom_point(size = 4, alpha = .7) +
  geom_point(size = 4, stroke = .4, shape = 1, color = "gray20")
```

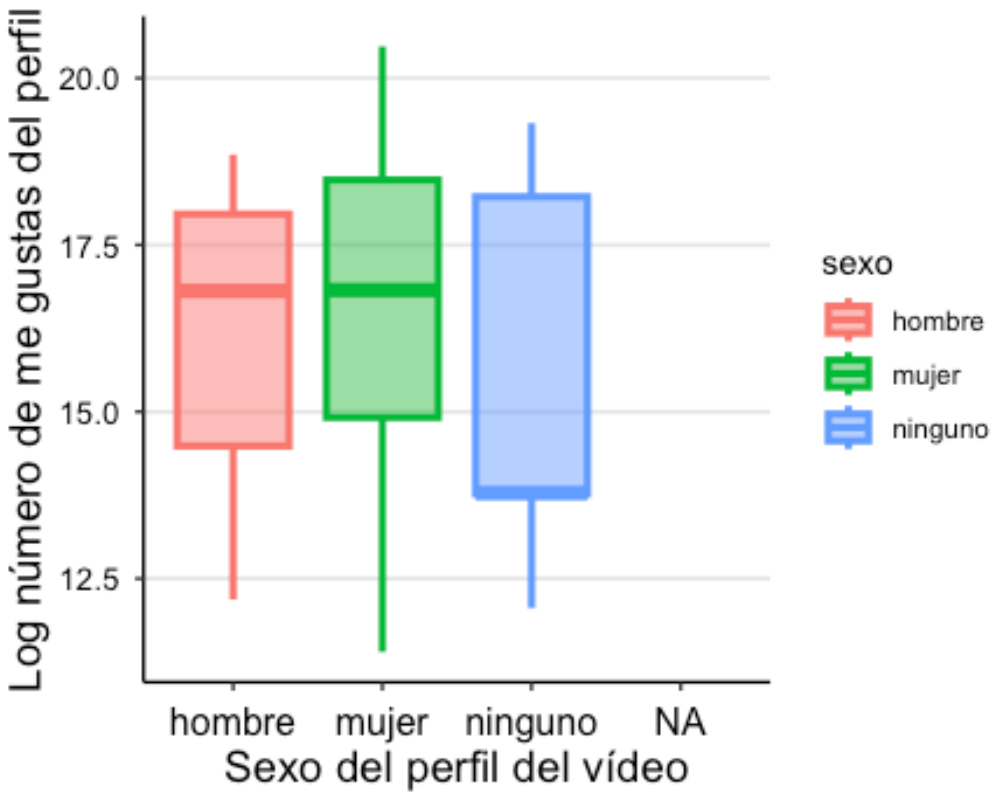



```
cor(log(tiktok$seguidores+1), log(tiktok$nmegustas+1), use="complete.obs")
```

```
## [1] 0.8116047
```

Sexo del perfil y número de me gustas del perfil

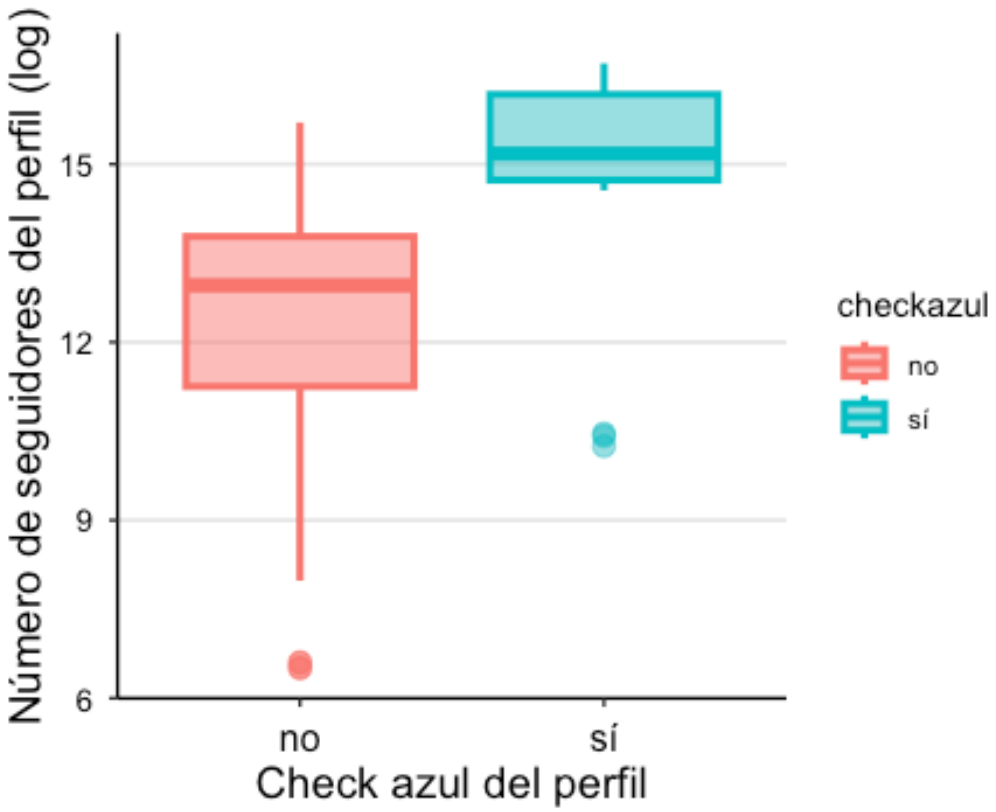
```
g4 <- ggplot(data=tiktok, aes(x = sexo, y = log(nmegustas), color = sexo, fill = sexo))
+
  scale_y_continuous() +
  labs(x="Sexo del perfil del vídeo",
       y="Log número de me gustas del perfil")

g4 +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```




```
tapply(tiktok$nmegustas , tiktok$sexo, summary)

## $hombre
##     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   196200   2000000  20000000  42350168  64750000 154100000
##
## $mujer
##     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    89500   3000000  20250000  95353091 105475000 780400000
##
## $ninguno
##     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##   172900    953400    966000  37892846  82250000 247100000         1

fit4 <- aov(lm(log(tiktok$nmegustas)~tiktok$sexo))
summary(fit4)

##              Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$sexo   2   31.4  15.697   2.875 0.0605 .
## Residuals   114  622.3   5.459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

Check azul y número de seguidores

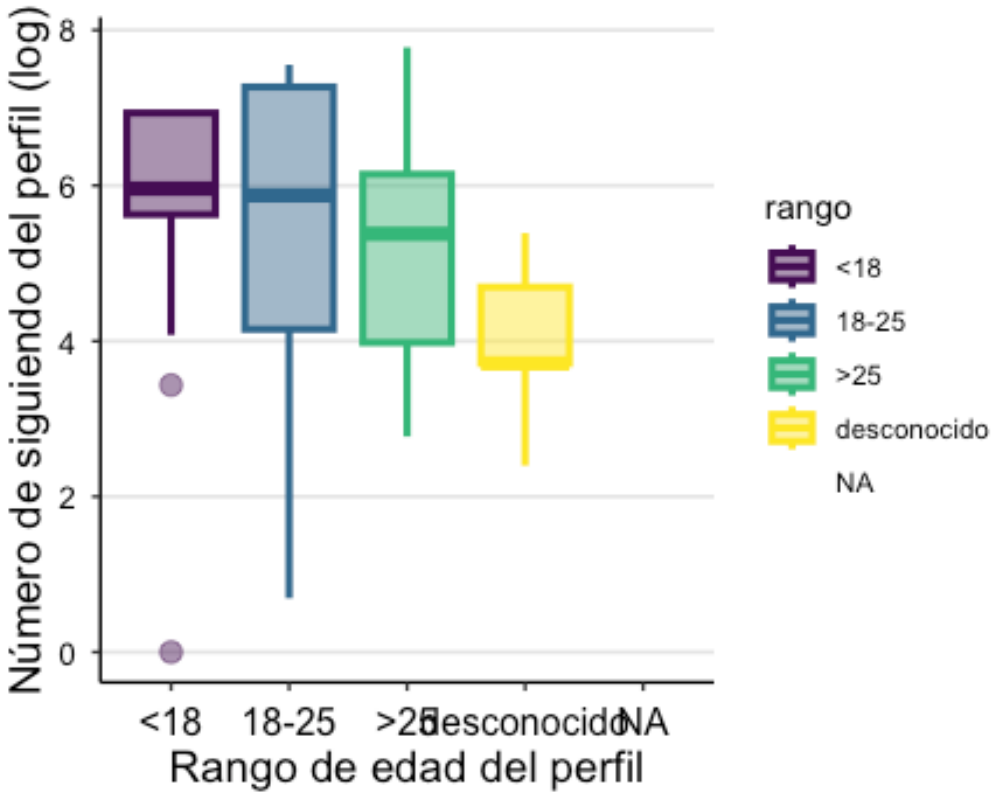
```
g5 <- ggplot(data=tiktok, aes(x = checkazul, y = log(seguidores), color = checkazul, fill
= checkazul)) +
  scale_y_continuous() +
  labs(x="Check azul del perfil",
       y="Número de seguidores del perfil (log)")

g5 +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```




```
tapply(tiktok$seguidores , tiktok$checkazul, summary)

## $no
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     670   77800  424200  911583  965950 6600000       3
##
## $sí
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   28200  2500000  3900000  6858932 10600000 17900000

fit5 <- aov(lm(log(tiktok$seguidores)~tiktok$checkazul))
summary(fit5)

##                   Df Sum Sq Mean Sq F value   Pr(>F)
## tiktok$checkazul   1  121.1  121.08   25.62 1.59e-06 ***
## Residuals        115  543.4    4.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

Rango de edad y los siguiendo del perfil

```
g6 <- ggplot(data=tiktok, aes(x = rango, y = log(siguiendo), color = rango, fill = rango))
+
  scale_y_continuous() +
  labs(x="Rango de edad del perfil",
       y="Número de siguiendo del perfil (log)")

g6 +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```



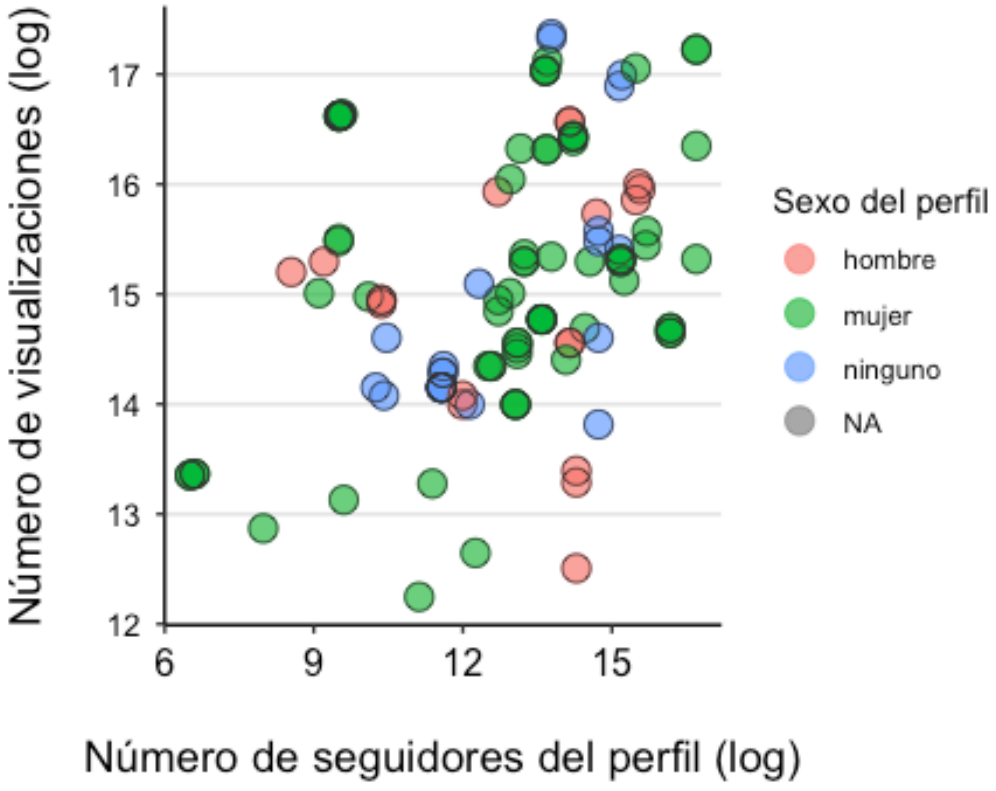

```
tapply(tiktok$siguiendo , tiktok$rango, summary)

## $`<18`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    23.5   312.5   413.2   780.8  1028.0
##
## $`18-25`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    69.0   356.0   673.8  1582.8  1908.0
##
## $`>25`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    16.0    53.5   217.0   471.7   467.8  2389.0       1
##
## $desconocido
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   41.00   41.00   69.64  115.75  219.00
```

```
##
## $`NA`
## NULL

fit6 <- aov(lm(log(tiktok$siguiendo+1)~tiktok$rango))
summary(fit6)

##              Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$rango   3   35.1  11.690   3.756  0.013 *
## Residuals    112  348.6   3.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 4 observations deleted due to missingness
```

3.2.4 Entre las métricas de las cuentas y de los vídeos

Visualizaciones del vídeo y seguidores de la cuenta

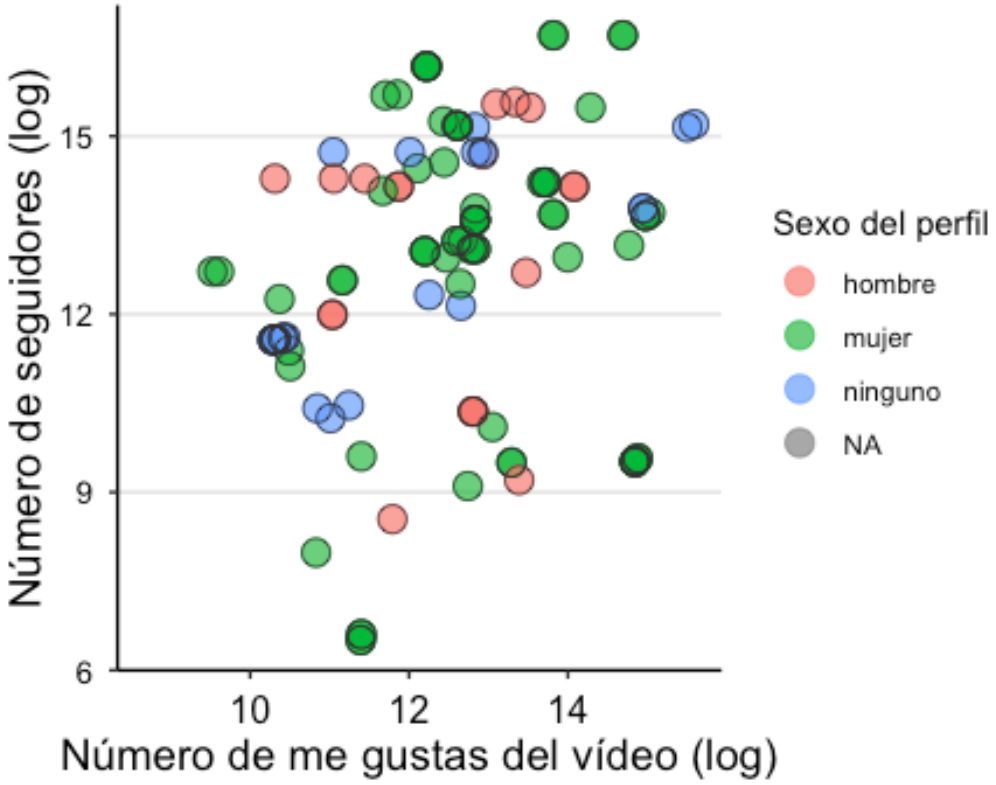
```
g <- ggplot(data=tiktok, aes(x = log(seguidores), y = log(viewsvid), color = sexo)) +
  labs(x="\nNúmero de seguidores del perfil (log)",
       y="Número de visualizaciones (log)\n",
       color = "Sexo del perfil")

# g+geom_point(size = 3, alpha = .7,  stroke = .4, shape = 1, color = "gray20")

g +
  geom_point(size = 4, alpha = .7) +
  geom_point(size = 4, stroke = .4, shape = 1, color = "gray20")
```



```
cor(log(tiktok$seguidores+1), log(tiktok$viewsvid+1), use="complete.obs")

## [1] 0.3060231
```

Me gustas del vídeo y seguidores del perfil

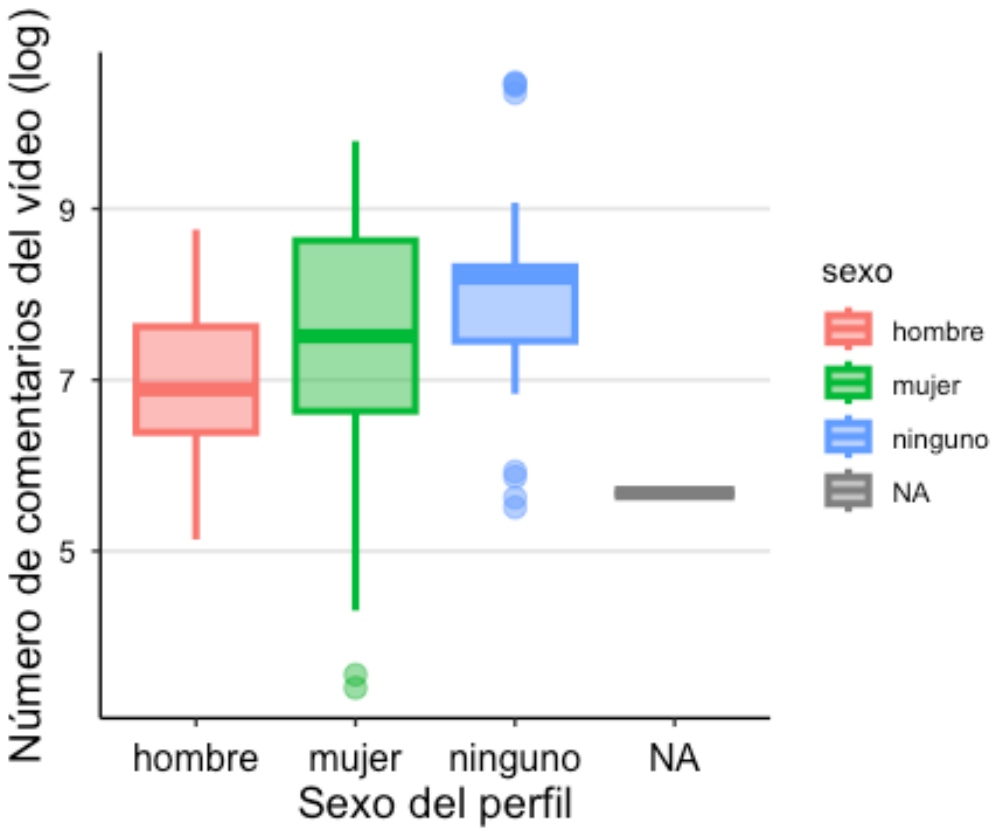
```
g7 <- ggplot(data=tiktok, aes(x = log(mgvid), y = log(seguidores), color = sexo)) +
  labs(x="Número de me gustas del vídeo (log)",
       y="Número de seguidores (log)",
       color = "Sexo del perfil")

# g+geom_point(size = 3, alpha = .7,  stroke = .4, shape = 1, color = "gray20")

g7 +
  geom_point(size = 4, alpha = .7) +
  geom_point(size = 4, stroke = .4, shape = 1, color = "gray20")
```




```
cor(log(tiktok$mgvid+1), log(tiktok$seguidores+1), use="complete.obs")

## [1] 0.1949263
```

Sexo del perfil y los comentarios del vídeo

```
g8 <- ggplot(data=tiktok, aes(x = sexo, y = log(comentarios), color = sexo, fill = sexo))
+
  scale_y_continuous() +
  labs(x="Sexo del perfil",
       y="Número de comentarios del vídeo (log)")

g8 +
  geom_boxplot(alpha = .5, size = 1.1, outlier.size = 3)
```




```
tapply(tiktok$comentarios , tiktok$sexo, summary)

## $hombre
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   170.0   620.5   984.0  1652.3  2043.5  6355.0
##
## $mujer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    30.0   767.2  1829.0  4296.6  5596.0 17900.0
##
## $ninguno
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     247    1727    3628    8057    4118   35500

fit7 <- aov(lm(log(tiktok$comentarios)~tiktok$sexo))
summary(fit7)

##              Df Sum Sq Mean Sq F value Pr(>F)
## tiktok$sexo   2   13.1   6.552   3.212 0.0439 *
## Residuals   115  234.6   2.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

4. MODELOS PREDICTIVOS

4.1. Modelo de regresión logística

```
mod3 <- glm(ctarec ~ sexo+siguiendo+sexo:siguiendo , data=ctas, family=binomial)
summary(mod3)

##
## Call:
## glm(formula = ctarec ~ sexo + siguiendo + sexo:siguiendo, family = binomial,
##     data = ctas)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8199  -0.9885   0.1584   1.0575   1.4591
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.261449   1.113468   1.133   0.2573
## sexomujer           -1.787035   1.224326  -1.460   0.1444
## sexoninguno          1.444905   2.071502   0.698   0.4855
## siguiendo           -0.007463   0.005104  -1.462   0.1437
## sexomujer:siguiendo  0.009514   0.005277   1.803   0.0714 .
## sexoninguno:siguiendo -0.015910   0.016519  -0.963   0.3355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 73.455  on 52  degrees of freedom
## Residual deviance: 59.255  on 47  degrees of freedom
## AIC: 71.255
##
## Number of Fisher Scoring iterations: 6
```

```
contrasts(ctas$ctarec)

##    si
## no  0
## si  1
```

```
probs_mod3 <- predict(mod3 , type = "response")
# Comprobamos que los valores obtenidos están comprendidos entre 0 y 1
summary(probs_mod3)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000821 0.379241 0.512690 0.509434 0.673834 0.987537
```

```
preds_mod3 <- rep("no", nrow(ctas))
preds_mod3[probs_mod3 > .5] <- "si"
```

```r
MCmod3<-table(preds_mod3 , ctas$ctarec)
MCmod3

##
## preds_mod3 no si
##         no 16  9
##         si 10 18
```

```r
(16 + 18) / (16+9+10+18)

## [1] 0.6415094
```

```r
mean(preds_mod3 == ctas$ctarec)

## [1] 0.6415094
```

```r
errores <- numeric(nrow(ctas))

# Guardamos en este vector las predicciones obtenidas mediante LOOCV
predLOOV <- character(nrow(ctas))

for( i in 1:nrow(ctas) ){
    # ajuste sin el la cuenta i
    lr.fit <- glm(ctarec ~ sexo+siguiendo+sexo:siguiendo, data=ctas[-i, ], family=binomial
)
    # probabilidad estimada para la cuenta i
    # utilizamos las columnas 1=sexo, y 3=siguiendo (son las var. explicativas del model
o)
    p.i<-predict(lr.fit, ctas[i, c(1,3)], type="response")
    # clasificamos esta cuenta según su probabilidad
    pred.i <- ifelse(p.i>0.5, "si", "no")
    predLOOV[i] <- pred.i
    # Vemos si el modelo la ha clasificado correctamente
    errores[i] <- ifelse(pred.i==ctas[i, 6], 0, 1) # columna 6=ctarec (var respuesta)
}

mean(errores)

## [1] 0.3773585
```

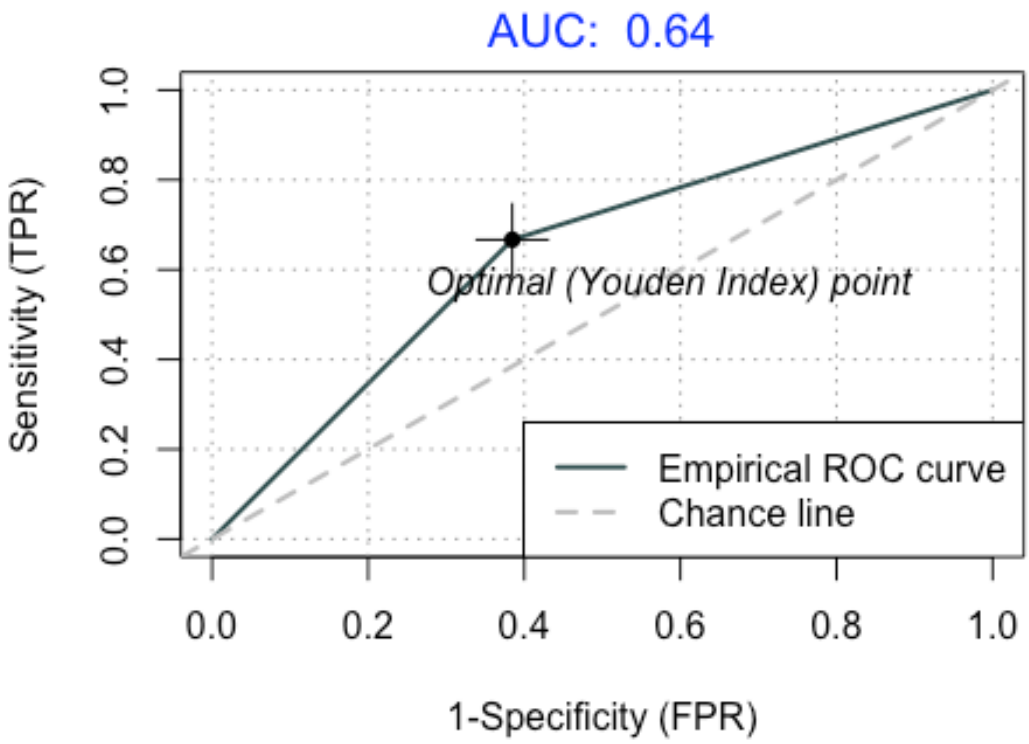
```r
metricas_de_error <- function(CM)
  # CM es la tabla con la matriz de confusión
{
  TN =CM[1,1] # True Negatives
  TP =CM[2,2] # True Positives
  FP =CM[1,2] # False Positives
  FN =CM[2,1] # False Negatives
```

```
  precision  = (TP+TN)/(TP+TN+FP+FN)


  tasaFalsosPositivos =(FP)/(FP+TN)
  tasaFalsosNegativos =(FN)/(FN+TP)

  print(paste("Precisión total del modelo: ",round(precision,2)))
  print(paste("Tasa de error: ", round(1-precision, 2)))

  print(paste("Tasa de falsos positivos: ",round(tasaFalsosPositivos,2)))
  print(paste("Tasa de falsos negativos: ",round(tasaFalsosNegativos,2)))
}
```

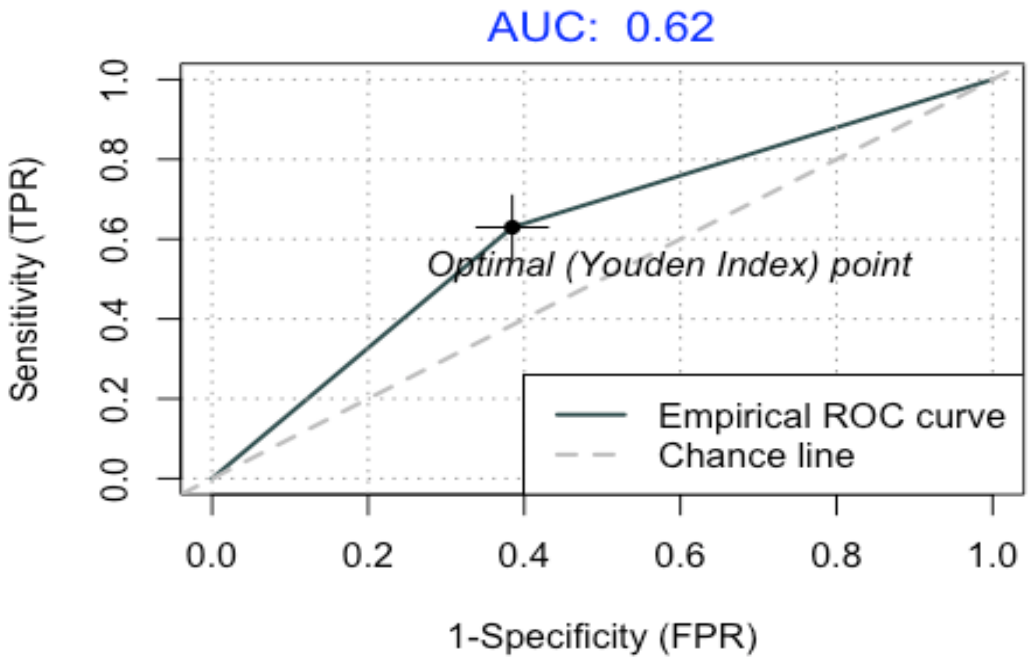
```
metricas_de_error(MCmod3)

## [1] "Precisión total del modelo:  0.64"
## [1] "Tasa de error:  0.36"
## [1] "Tasa de falsos positivos:  0.36"
## [1] "Tasa de falsos negativos:  0.36"
```

```
MCLOOCV <- table(predLOOV, ctas$ctarec)
metricas_de_error(MCLOOCV)

## [1] "Precisión total del modelo:  0.62"
## [1] "Tasa de error:  0.38"
## [1] "Tasa de falsos positivos:  0.38"
## [1] "Tasa de falsos negativos:  0.37"
```

```
# necesitamos 1's y 0's en lugar de "si" y "no" para hacer esta función del paquete RO
Cit

vpredics <- ifelse(preds_mod3=="si", 1, 0)

library(ROCit)

ROCrl <- rocit(score=vpredics, class=ctas$ctarec)
plot(ROCrl)
mtext(paste("AUC: ", round(ROCrl$AUC,2)), side=3, padj=-0.5, col="blue", cex=1.25)
```

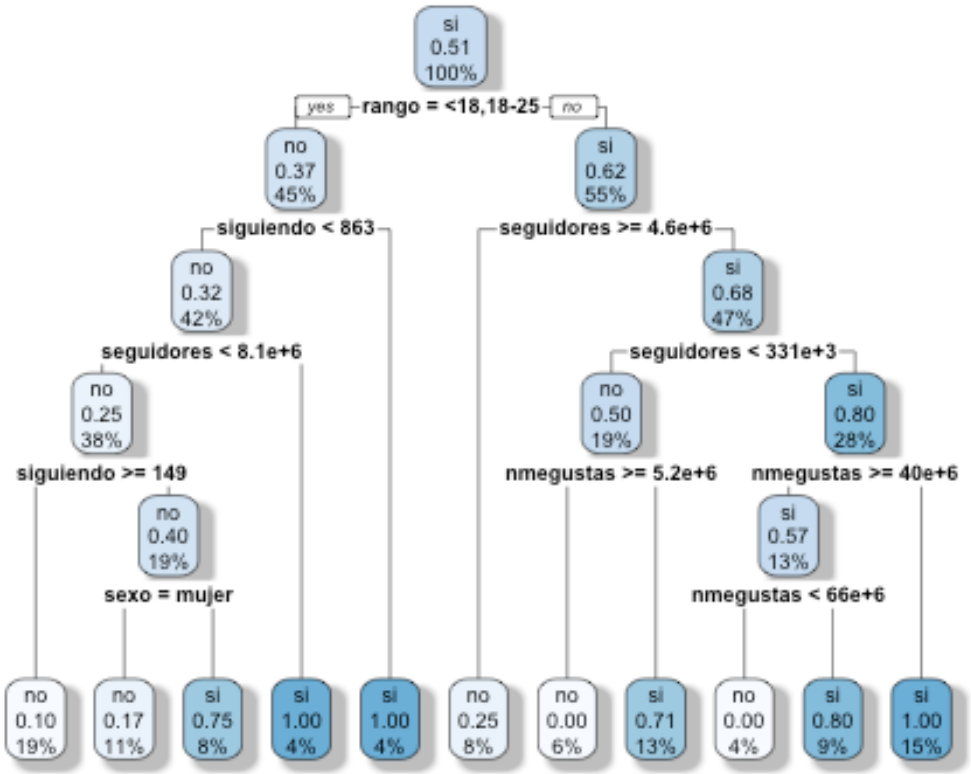



```
# necesitamos 1's y 0's en lugar de "si" y "no" para hacer esta función del paquete RO
Cit

vpredics <- ifelse(predLOOV=="si", 1, 0)

library(ROCit)

ROCrlLOO <- rocit(score=vpredics, class=ctas$ctarec)
plot(ROCrlLOO)
mtext(paste("AUC: ", round(ROCrlLOO$AUC,2)), side=3, padj=-0.5, col="blue", cex=
1.25)
```



4.2 Arboles de decisión

```
ctas$perfil <- NULL
# load the package
library(rpart)

# fit model
arbol1 <- rpart(ctarec ~ . , data=ctas, control=rpart.control(minsplit=5))
# summarize the fit
# summary(arbol1)

# Obtain  a picture of the decision tree
library(rpart.plot)
# You can set plotting parameters to change colour, shape, and information displayed (
which and where) on the nodes.

rpart.plot(arbol1, type = 2, extra = "auto",
  box.palette="Blues", shadow.col="gray")
```

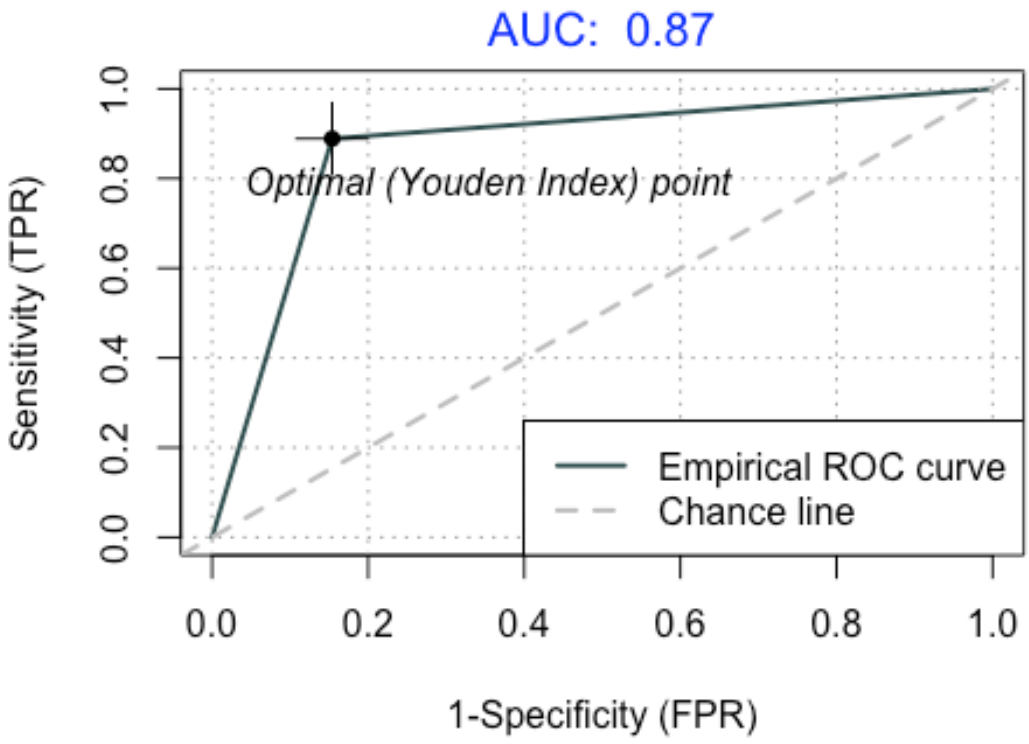


```r
# 1. Probabilidades calculadas por el árbol
probsArbol1 <- predict(arbol1)[,"si"]
# summary(probsArbol1)

# 2. Asignamos la clase "si" para prob>0.5, y "no" para el resto
# El punto de corte 0.5 es arbitrario, pero es el valor lógico en este caso
predArbol1 <- ifelse(probsArbol1>0.5, "si", "no")

# 3. Matriz de confusión
MCArbol1 <- table(predArbol1, ctas$ctarec)
MCArbol1

##
## predArbol1 no si
##         no 22  3
##         si  4 24

# Métricas de error
metricas_de_error(MCArbol1)

## [1] "Precisión total del modelo:  0.87"
## [1] "Tasa de error:  0.13"
## [1] "Tasa de falsos positivos:  0.12"
## [1] "Tasa de falsos negativos:  0.14"
```

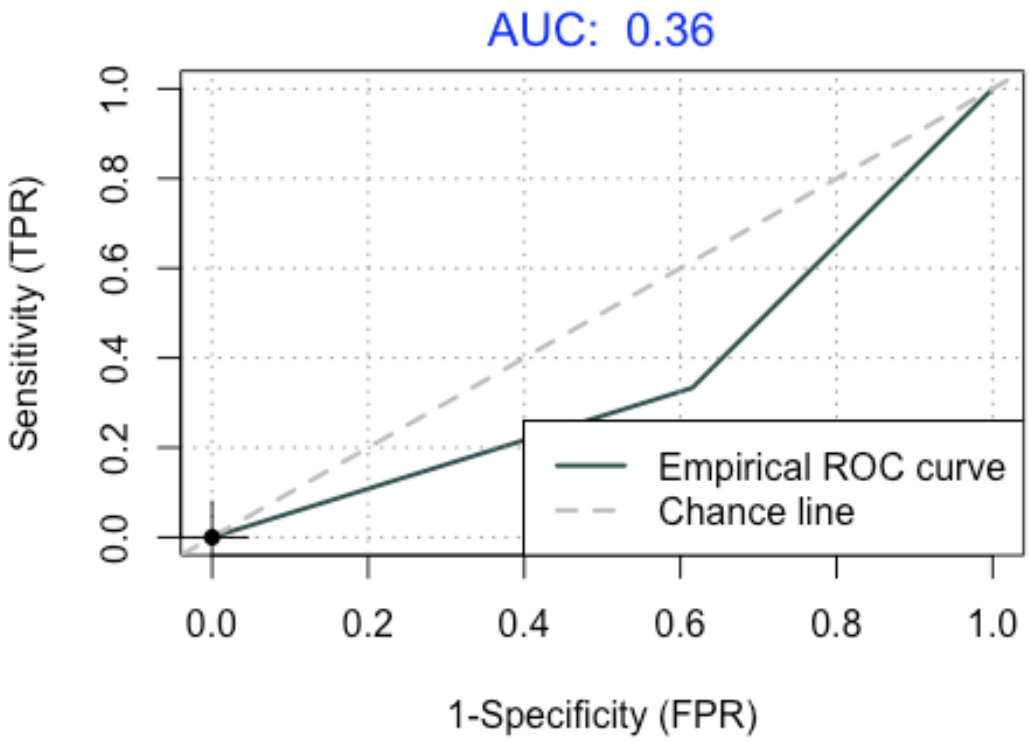
```r
errores <- numeric(nrow(ctas))

# Guardamos también las predicciones obtenidas mediante LOOCV
predLOO <- character(nrow(ctas))

for( i in 1:nrow(ctas) ){
   # ajuste sin la cuenta i
   arbol <- rpart(ctarec ~ . , data=ctas[-i, ], control=rpart.control(minsplit=5))

   # probabilidad estimada para la cuenta i
   p.i<-predict(arbol, ctas[i,])["si"]

   # clasificamos esta cuenta según su probabilidad
   pred.i <- ifelse(p.i>0.5, "si", "no")
   predLOO[i]<-pred.i

   # Vemos si el modelo la ha clasificado correctamente
   errores[i] <- ifelse(pred.i==ctas[i, 6], 0, 1) # columna 6=ctarec (var respuesta)
}

mean(errores)

## [1] 0.6415094
```

```
MCarbolLOO <- table(predLOO, ctas$ctarec)
metricas_de_error(MCarbolLOO)

## [1] "Precisión total del modelo:  0.36"
## [1] "Tasa de error:  0.64"
## [1] "Tasa de falsos positivos:  0.64"
## [1] "Tasa de falsos negativos:  0.64"
```

```
# necesitamos 1's y 0's en lugar de "si" y "no" para hacer esta función del paquete RO
Cit

vpredics <- ifelse(predArbol1 =="si", 1, 0)

library(ROCit)

ROCarbol <- rocit(score=vpredics, class=ctas$ctarec)
plot(ROCarbol)
mtext(paste("AUC: ", round(ROCarbol$AUC,2)), side=3, padj=-0.5, col="blue", cex=1.
25)
```

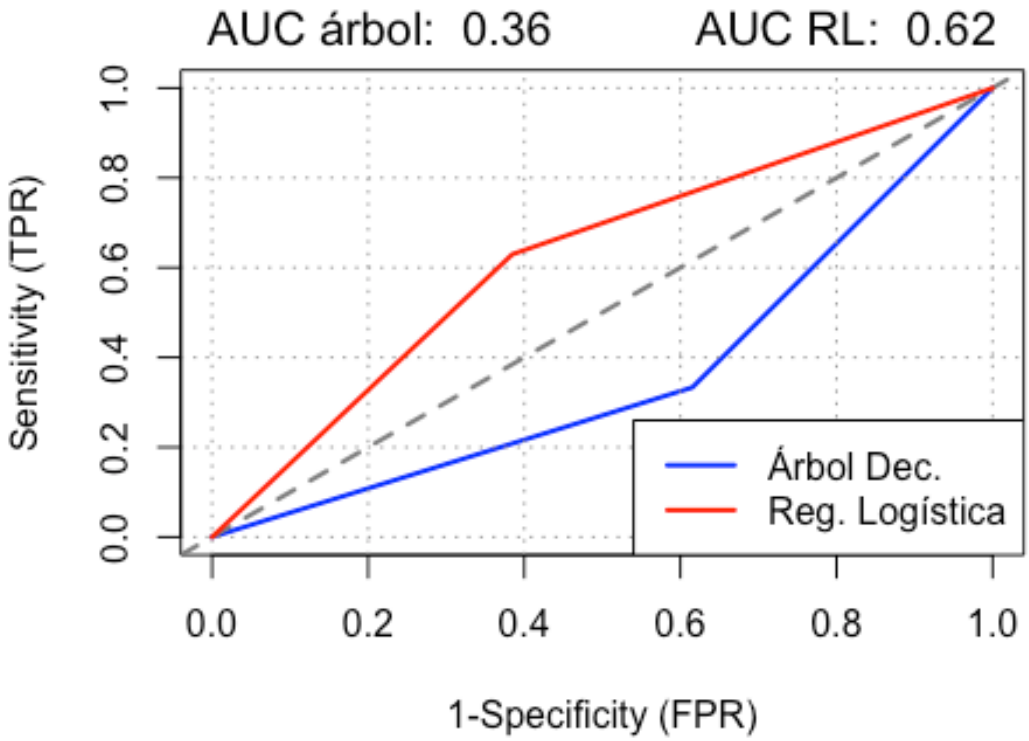


```
# necesitamos 1's y 0's en lugar de "si" y "no" para hacer esta función del paquete RO
Cit

vpredics <- ifelse(predLOO=="si", 1, 0)

library(ROCit)

ROCarLOO <- rocit(score=vpredics, class=ctas$ctarec)
plot(ROCarLOO)
mtext(paste("AUC: ", round(ROCarLOO$AUC,2)), side=3, padj=-0.5, col="blue", cex=
1.25)
```



4.3. Comparación de modelos

```
plot(ROCarLOO, col = c("blue","gray50"), legend = FALSE, YIndex = FALSE)
lines(ROCrlLOO$TPR~ROCrlLOO$FPR, col = "red", lwd = 2)
legend("bottomright", col = c("blue","red"),
       c("Árbol Dec.", "Reg. Logística"), lwd = 2)
mtext(paste("AUC árbol: ", round(ROCarLOO$AUC,2), "          AUC RL: ", round(RO
CrlLOO$AUC,2)), side=3, padj=-0.5, cex=1.25)
```


AUC árbol: 0.36
AUC RL: 0.62
Sensitivity (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity (FPR)
Árbol Dec.
Reg. Logística