Escola Técnica Superior de Enxeñaría

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Visualización Avanzada

Universidad de Santiago de Compostela

Práctica 5.8: TikTok: vídeos populares, autores y estadísticas

- Hugo Vázquez Docampo

FECHA DE ENTREGA: 12 de Diciembre de 2022

´

Índice

1. Introducción **2**

2. Motivación, preguntas a resolver y origen de los datos **2**
A. Fuente de datos 2

3. Diccionario de datos **3**

4. Preparación y limpieza de los datos **4**

5. Elaboración del Dahboard 1: ranking de usuarios **4**

6. Dahboard 2: influencia de la verificación en la popularidad **5**

7. Dahboard 3: otros análisis de interés **7**

8. Campos calculados definidos **7**

9. Análisis de los resultados observados **8**

10. Conclusiones **9**

1. Introducción

Se incluye en este informe un análisis de un conjunto de datos de libre elección sobre **métricas** asociadas a distintas propiedades y componentes de TikTok. En él, se incluye un análisis de las cuestiones que se desean resolver; así como una motivación para la selección de este conjunto de datos. Además, se incorpora una breve descripción de los datos mostrados. A mayores, se introduce un apartado en el cual se relatan los pasos a seguir para realizar la limpieza de los datos. A continuación, se lleva a cabo el análisis de los mismos mediante la realización de una serie de Dashboards. Por último, se finaliza con una conclusión de los resultados observados.

2. Motivación, preguntas a resolver y origen de los datos

TiktTok es una plataforma de origen china que basa su experiencia de uso en la muestra de videos cortos (de duración variable entre 15 segundos y 3 minutos) sobre diferentes temas. El factor diferencial de esta herramienta se basa en su potente algoritmo de recomendación que permite a los usuarios obtener siempre el contenido que mejor se adapte a sus gustos.

Además, el hecho de permitir a cualquier usuario ser creador de contenido aumenta considerablemente las posibilidades de uso de la aplicación. Cada persona es libre para dejar fluir su creatividad y mostrar el contenido que mejor se adapte a su forma de ser. Por ello, se considera de interés realizar un análisis que permita conocer más acerca de los principales **creadores de contenido** a nivel global y, con ayuda de las variables disponibles en el dataset, tratar de inferir cual es la clave de su popularidad.

Para poder realizar este análisis, se plantean las siguientes cuestiones:

1. ¿Cómo influye el número de likes y de reproducciones en la popularidad de los usuarios?
2. ¿Los usuarios verificados son los más populares? ¿Por qué?
3. ¿La duración de los vídeos es un factor fundamental en la popularidad de los vídeos?
4. Los videos más compartidos, ¿a qué usuarios pertenecen?

A. Fuente de datos

Los datos utilizados para la realización de este análisis se han obtenido de la siguiente fuente: Popular TikTok Videos, Authors, and Musics. Se trata de un dataset de la plataforma Kaggle con información acerca del usuario que publicó cada video, la cantidad de likes, shares y comentarios de cada publicación y otra información de interés como la duración de los vídeos.

3. Diccionario de datos

Se recoge a continuación un diccionario de datos asociado a las columnas de los conjuntos de datos utilizados para el análisis:

Tabla Trending_authors:

- **Author Nickname:** nombre del usuario dentro de la plataforma. Tipo:String.
- **Author Unique ID:** id del usuario en la aplicación. Tipo: String.
- **Avatar Thumbnail:** imagen u avatar empleado por el usuario. Tipo: String. Este campo no se puede emplear por no disponer de permisos para acceder al contenido.
- **Signature:** firma asociada al usuario. Tipo: String.
- **Verified?:** campo que indica si el usuario está o no verificado o no en la aplicación. Tipo: Booleano. Por problemas a la hora de llevar a cabo la representación de este campo, ha sido necesario definir un KPI llamado "Verificado" con la siguiente fórmula:

  ```
  IIF([Verified?]=True,'Verdadero','Falso')
  ```

  Esta únicamente devuelve verdadero en caso de que el usuario esté verificado y falso en caso contrario.
- **Private Account?:** muestra si el usuario posee una cuenta privada o no en tiktok. Tipo: Booleano.

Tabla Trending_videos:

- **user_name:** nombre del usuario dentro de la plataforma. Tipo:String.
- **video_id:** identificador del vídeo en tiktok. Tipo: String.
- **video_desc:** descripción del vídeo. Tipo: String.
- **video_length:** duración del vídeo. Tipo: Entero.
- **video_link:** enlace para acceder al vídeo en la propia aplicación. Tipo: String.
- **n_likes:** número de likes que un vídeo ha recibido. Tipo: Entero.
- **n_shares:** número de veces que un vídeo ha sido compartido. Tipo: Entero.
- **n_comments:** cantidad de comentarios que un vídeo posee. Tipo: Entero.

tiktok_collected_liked_videos:

- **user_name:** nombre del usuario dentro de la plataforma. Tipo:String.
- **user_id:** identificador del usuario dentro de la plataforma. Tipo:String.
- **video_id:** identificador del vídeo en tiktok. Tipo: String.
- **video_desc:** descripción del vídeo. Tipo: String.
- **video_length:** duración del vídeo. Tipo: Entero.
- **video_link:** enlace para acceder al vídeo en la propia aplicación. Tipo: String.
- **n_likes:** número de likes que un vídeo ha recibido. Tipo: Entero.
- **n_shares:** número de veces que un vídeo ha sido compartido. Tipo: Entero.
- **n_comments:** cantidad de comentarios que un vídeo posee. Tipo: Entero.

tiktok_collected_liked_videos → Campos calculados:

- **Ranking likes:** se define un campo calculado que, mediante la función RANK, lleve a cabo un ranking del número de likes que posee cada usuario.
- **Ranking reproducciones:** se define un campo calculado que, mediante la función RANK, lleve a cabo un ranking del número de reproducciones que poseen, en conjunto, los vídeos asociados a cada usuario.

4. Preparación y limpieza de los datos

Antes de comenzar el proceso de análisis se han abierto los distintos ficheros csv asociados al dataset para comprobar la existencia de posibles valores atípicos y, en caso de encontrar algo que se pudiese solucionar, tratar de solventar los problemas.

De este modo, se aprecian una serie de problemas sobre el conjunto de datos. No obstante, estos no eran resolubles de forma razonable (o bien no afectan a los análisis que se desean realizar) y, por ello, se han empleado los ficheros originales descargados del sitio web. Los errores detectados han sido:

- El link a alguno de los vídeos estaban caídos; posiblemente porque el autor o la propia compañía ha considerado prudente eliminar ese contenido.
- Los avatares asociados a los diferentes usuarios no poseían permisos de lectura por lo que no ha sido posible incluirlos en ninguno de los dashboards elaborados.

Tras comprobar la validez de los datos, se procede a realizar la carga de los mismos en Tableau. Así, una vez abierta la aplicación se escoge la opción de **conectar a un archivo de texto** y se selecciona el fichero de "tiktok_collected_liked_videos.csv"; tras esto, se arrastra al espacio en blanco del centro de la pantalla. Esta fuente de datos ya es usable.

A continuación, en el menú superior, en la sección de **datos**, se selecciona **nueva fuente de datos** y se introducen los archivos "Trending_authors.csv" y Trending_videos.csv". Se arrastra el primero de ellos al centro de la pantalla y, a continuación, el segundo. Una vez realizado esto, es necesario configurar la conexión de ambos ficheros. Para ello, se hace click sobre la línea que los conecta y se escogen como columans de conexión **Author Unique ID** (de Trending_authors) y **user_name** de (Trending_videos). A continuación, pulsando con el botón derecho sobre cada archivo, se define en **propiedades del archivo de texto** que el delimitador a utilizar es una coma. Con esto, ya es posible empezar a realizar las diferentes gráficas.

5. Elaboración del Dahboard 1: ranking de usuarios

El primero de los Dashboards consta de dos tablas que conforman un ranking de la popularidad de los usuarios en función de su número de likes y reproducciones. Además, se incluye en la parte inferior del mismo un gráfico de segmentación de datos para poder filtrar en función de la duración de los vídeos (dado que este es uno de los aspectos cuya influencia se busca determinar).

Así, para la elaboración de estas tablas se arrastra el campo **User name** de la tabla "tiktok_collected_liked_videos" a las filas. A continuación, se mueven los campos **N Likes / N Plays** y **ranking likes / ranking reproducciones** a las columnas. Después, desde el menú de muéstrame, se escoge la opción **tablas de texto**. Por último, se pulsa sobre la cabecera de la columna asociada al ranking y, una vez esté toda la columna seleccionada, en el menú superior se selecciona el icono de **ordenar User name de forma ascendente en función de los valores de las medidas.** Por último, se guardan las hojas de trabajo con el nombre deseado.

Tras haber creado ambas tablas, es el momento de elaborar el Dashboard en el que se recogerán ambas representaciones. Para ello, se pulsa en el botón de **Nuevo DashBoard** y se arrastran las gráficas deseadas sobre la posición que se desea establecer en el informe.

Una vez acopladas sobre el mismo, sólo faltaría introducir el filtro de segmentación. Para ello, sobre la hoja asociada a una de las tablas, se arrastra el campo **Video Length** y se escoge

como medida el **promedio**. Tras hacer esto, en el DashBoard creado se pulsa con el botón derecho en la parte inferior izquierda de la gráfica asociada a la hoja sobre la que se aplicó el filtro y, en filtro, se escoge el que acabamos de introducir.

Tras ubicarlo en el lugar deseado, solo faltaría conectarlo con la otra de las tablas. Para ello, pulsando con el botón derecho sobre dicho filtro, se escoge la opción de **aplicar a hojas de trabajo >Hojas de trabajo seleccionadas.** Entonces, se seleccionan las de las dos tablas y se guarda el trabajo.

Una vez finalizado este proceso, el resultado obtenido es el siguiente:

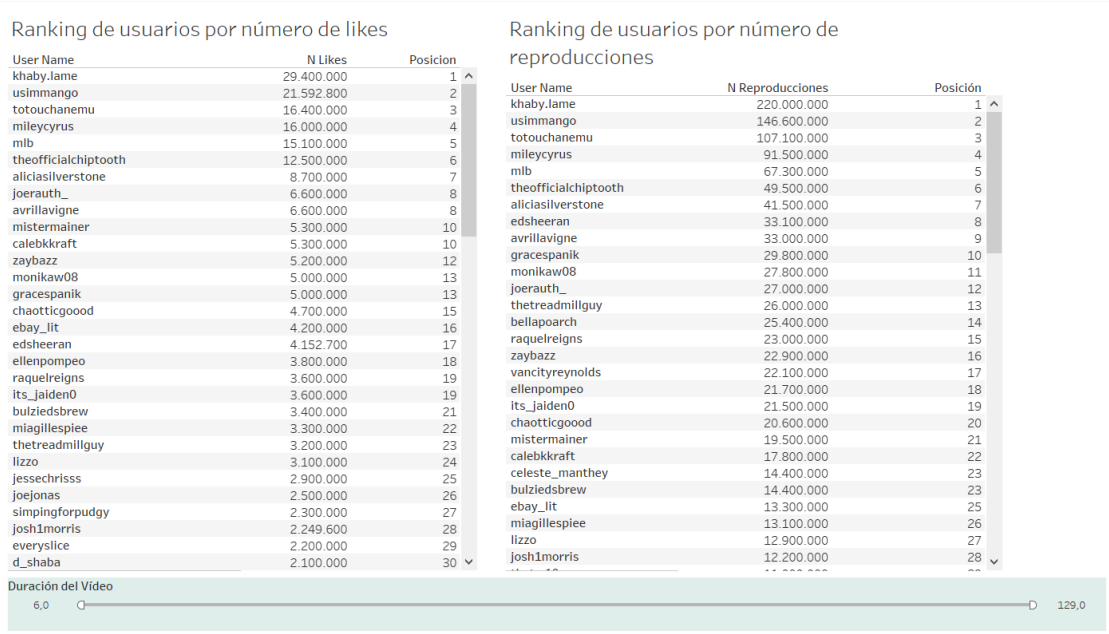

Figura 1: *Dashboard 1:ranking de usuarios*

Cabe mencionar que, la explicación y el código asociados a los campos calculados se definirá en la sección correspondiente del informe.

6. Dahboard 2: influencia de la verificación en la popularidad

Con este informe se busca responder a las preguntas asociadas a la influencia de la verificación de las cuentas de usuario sobre la popularidad de los mismos. Para ello, se han definido dos representaciones; configurando una de ellas para que sirva como gráfica de segmentación y permita filtrar los datos.

En primer lugar se definirá un **gráfico circular** que indique el porcentaje de usuarios verificados. Para ello, se crea una nueva hoja de trabajo y, en el menú de **marcas**, se selecciona la opción de **Circular**.

Tras esto, se arrastra el campo **user_name** sobre el ángulo y se cambia la configuración de Dimensión a medida >**recuento** y se escoge en **cálculo de tablas rápido** porcentaje del total. Después, se introduce el campo calculado ”Verificado” en la opción de **color** sobre **marcas**.

A continuación, para dar formato a la representación, se aumenta el tamaño de la misma seleccionando **Vista Completa** en el menú desplegable superior y se añaden las etiquetas de datos desde el menú de marcas >**etiquetas** >**Mostrar etiquetas de marcas**. Por último, es necesario eliminar aquellas entradas asociadas a los valores nulos. Para ello se arrastra el campo calculado anteriormente utilizado a los filtros y se selecciona únicamente los valores verdadero y falso. Entonces, se cambia el nombre de la hoja y se guarda el trabajo.

Para la segunda de las representaciones, se crea una nueva hoja de trabajo y se arrastra el campo **Author nickname** de la tabla Trending_authors" a las filas y los campos **n_likes** y **n_plays** a las columnas (estableciendo como medida suma). Después se selecciona en **Mostrar-me**, tabla de texto.

A continuación, se arrastra el campo "verificado" sobre el color en el menú de **marcas** e, igual que antes, se filtran aquellas entradas con valores nulos. Entonces, se arrastra el campo **n_likes** a los filtros y se escoge suma y, en especial, **valores no NULL**. Se filtra también el usuario null del mismo modo con la variable Author nickname.

Por último, haciendo click sobre la esquina derecha de la columna del número de reproducciones, se ordena la tabla en función de este campo y se guarda el trabajo.

Una vez creadas las dos representaciones asociadas a este segundo Dashboard, se procede a crear el mismo. Para ello, igual que en el caso anterior se arrastran las gráficas al lugar deseado. Por último, solo queda incluir la representación del **% de usuarios verificados** como un filtro. Esto se realiza pulsando con el botón derecho sobre la representación y seleccionando la opción de **Usar como filtro**.

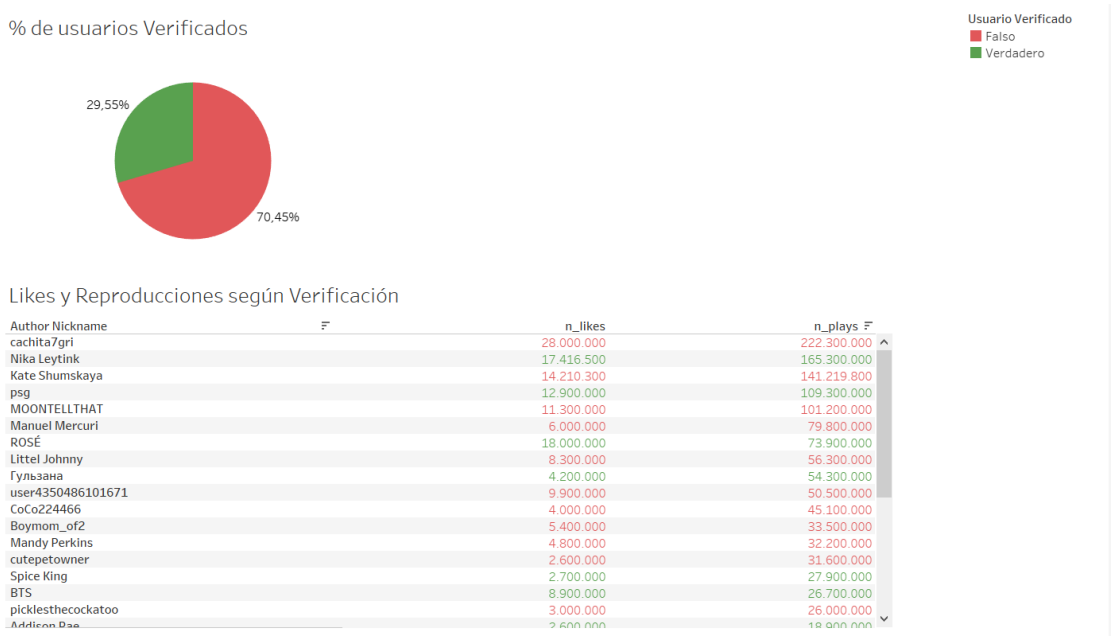

Figura 2: *Dashboard 2: influencia de la verificación en la popularidad*

7. Dahboard 3: otros análisis de interés

Este último Dashboard constará de dos nuevas gráficas y se incluirá además el filtro de segmentación creado para el primer informe. Por ello, el proceso de creación del mismo no se recogerá en este apartado dado que es igual que para el caso anterior.

La primera de las gráficas que se introducirá será una gráfica de burbujas que represente la influencia de la duración de los vídeos en el número de likes y de reproducciones. Para ello, se arrastra el campo **video length** a las columnas y el **N Likes** a las filas.

A continuación, desde **Mostrarme**, se escoge **diagrama de dispersión**. Por último, se establece en el menú de **marcas**, el nombre de usuario como color y el número de reproducciones al tamaño. Finalmente, en el menú de tamaño se mueve este hasta la mitad del tamaño posible y se guarda la hoja de trabajo.

Para la otra de las gráficas que conforma este informe se arrastra el campo **user name** a las filas y el **N Shares**, como suma, a las columnas. A continuación, se ordenan los datos en función del valor de las filas (número de veces compartidas).

Por último, se le aplica formato a la misma cambiando su color, añadiendo las etiquetas de datos (desde el menú de marcas) e introduciendo el recuento de vídeos en una descripción emergente (mediante el campo video id). Finalmente, se guarda la gráfica y se configura el Dashboard deseado.

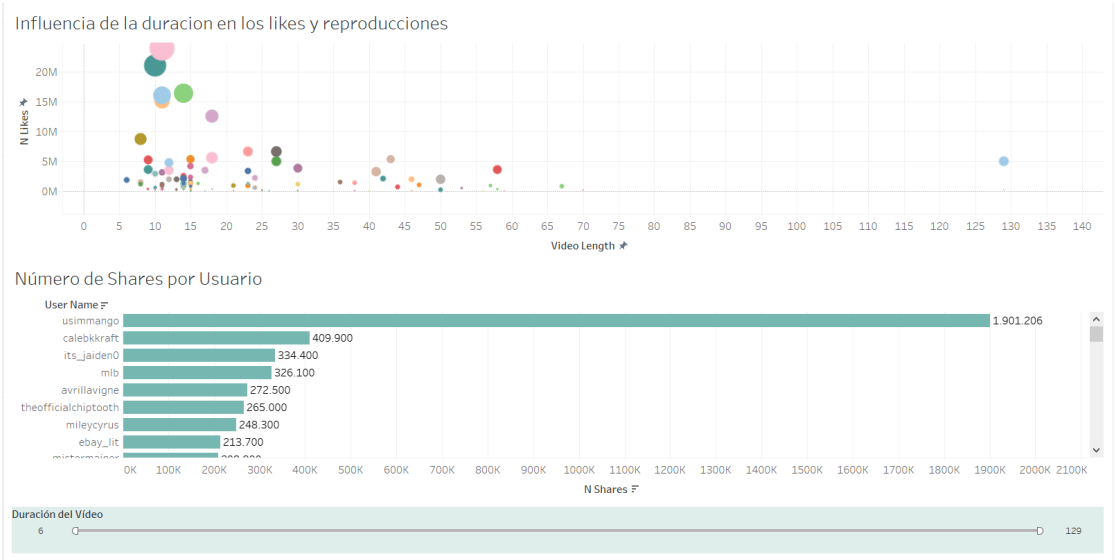

Figura 3: *Dashboard 3: otros análisis de interés*

8. Campos calculados definidos

Tal como se ha mencionado con anterioridad, para poder realizar algunas de las representaciones anteriormente incluidas, ha sido necesario definir una serie de campos calculados. Estos son:

- **verificado:** este campo se ha explicado en el diccionario de datos y busca poder utilizar el atributo **verified?** para ciertas representaciones mediante el siguiente código:

```
IIF([Verified?]=True,'Verdadero','Falso')
```

- **ranking likes:** este campo hace uso de la función rank para establecer un ranking con el número de likes para poder ordenar a los usuarios. El código asociado al mismo es el siguiente:

```
RANK(sum([N Likes]),'desc')
```

- **ranking reproducciones:** su uso es similar al del ranking de likes pero asociado a la variable de reproducciones:

```
RANK(sum([N Plays]),'desc')
```

9. Análisis de los resultados observados

Para justificar los resultados observados, se utiliza como punto de partida las preguntas planteadas al inicio del informe. Así, empleando como material de apoyo los dashboard elaborados, se recogen los siguientes resultados:

1. **¿Cómo influye el número de likes y de reproducciones en la popularidad de los usuarios?**
   Para dar respuesta a esta pregunta se hará uso del **Dashboard 1**. Como se puede observar, los usuarios que poseen más likes suelen ser también los mismos que ocupan los lugares más altos en el ranking de reproducciones. Por ello, es posible confirmar que ambas variables están muy relacionadas y, juntas, definen la popularidad de un usuario en esta plataforma. No obstante, se puede observar como hay ligeras variaciones.

   Por ejemplo, el cantante Ed Sheeran, ocupa el puesto número 8 en el ranking de número de reproducciones mientras que se encuentra en la posición número 17 del número de likes. Lo mismo sucede con otros artistas dedicados al mundo de la música. Esto tiene sentido porque la relación entre el número de likes y de reproducciones no es 1 a 1 y, muchas veces, si una canción incluida en un vídeo es del gusto del usuario, la reproducirá en más de una ocasión.

   Por ello, es posible concluir que, aunque ambos factores influyen directamente en la fama de los usuarios, no es posible obtener una regla única para su relación con la influencia de los creadores de contenido, pues depende de otros factores como el tipo de contenido publicado.

2. **¿Los usuarios verificados son los más populares? ¿Por qué?**
   Con ayuda del **Dashboard 2** es posible deducir que no pues, muchos de los usuarios que ocupan los puestos más altos en la lista de popularidad no poseen esta cualidad.

   Este indicativo únicamente muestra la veracidad de una cuenta para demostrar así que el usuario que publica el contenido es fiable. No obstante, esta falsa creencia está muy extendida porque normalmente, solo los personajes públicos disponen de esta validación.

   Sin embargo, como se comentó al inicio del informe, todos los usuarios son libres de publicar su propio contenido y, una persona, puede llegar a ser viral gracias a un vídeo concreto sin ser conocido públicamente hasta dicho momento.

3. **¿La duración de los vídeos es un factor fundamental en la popularidad de los vídeos?**
   Dado que la popularidad no es una métrica deducible a partir de un solo campo, es necesario hacer uso de los **Dasboard 1 y 3** para ciertos detalles importantes. Fundamentalmente, la gráfica de dispersión es la que más información aporta. Como se puede observar claramente, los círculos de mayor tamaño y ubicados en la parte mas alta del eje Y (es decir, los de mayor número de likes y reproducciones) están asociados a aquellos vídeos de menor duración.

   Además, si contrastamos los valores asociados a los globos de mayor tamaño con las posiciones de los usuarios en los diferentes ranking del Dashboard 1, observamos como estos usuarios ocupan los puestos más altos de los mismos.

   Por otra parte, si en el **Dashboard 1** vamos reduciendo la duración progresivamente y observamos las posiciones de la clasificación, apreciamos como estos valores apenas varían. Atendiendo a todas estas evidencias podemos dedicir que los vídeos más cortos son los que causan mayor impacto en los usuarios y, por tanto, son los que mayor probabilidad tienen de volverse virales.

4. **Los vídeos más compartidos, ¿a qué usuarios pertenecen?**
   Por último, para responder a esta pregunta se hará uso de la última gráfica del Dashboard 3, en la que se puede observar cómo el usuario que mayor número de shares posee es **ussimmango**; el cual ocupa el segundo puesto en los ranking elaborados. Por ello, también es de interés comentar que este campo no está estrictamente relacionado con la popularidad de un usuario.

10. Conclusiones

Tras la realización de este informe es posible concluir que tiktok es una red social de gran importancia y con mucha influencia en la sociedad (deducible a partir de los altos valores asociados a las variables de reproducciones, comentarios, likes...)

Además, ha sido posible observar datos interesantes como que los usuarios de mayor fama son aquellos que poseen en sus vídeos un mayor número de likes y reproducciones (no influyendo tanto el número de veces que se comparten sus vídeos).

También se ha observado cómo, el hecho de que un usuario esté verificado no es un factor que influya en la popularidad de los mismos.

Así mismo, también se ha descubierto que los datos no son interpretables siguiendo una única regla; pudiendo observar cómo el número de reproducciones y likes difiere en función del contenido publicado por los diferentes creadores.

Por último se ha observado la relevancia de la duración de los videos en la popularidad de los usuarios, mostrando cómo aquellos que publican vídeos de menor duración son, por regla general, los que mayor popularidad poseen en el publico