Prueba práctica Científico de Datos

Ing. Kenny Miranda D.

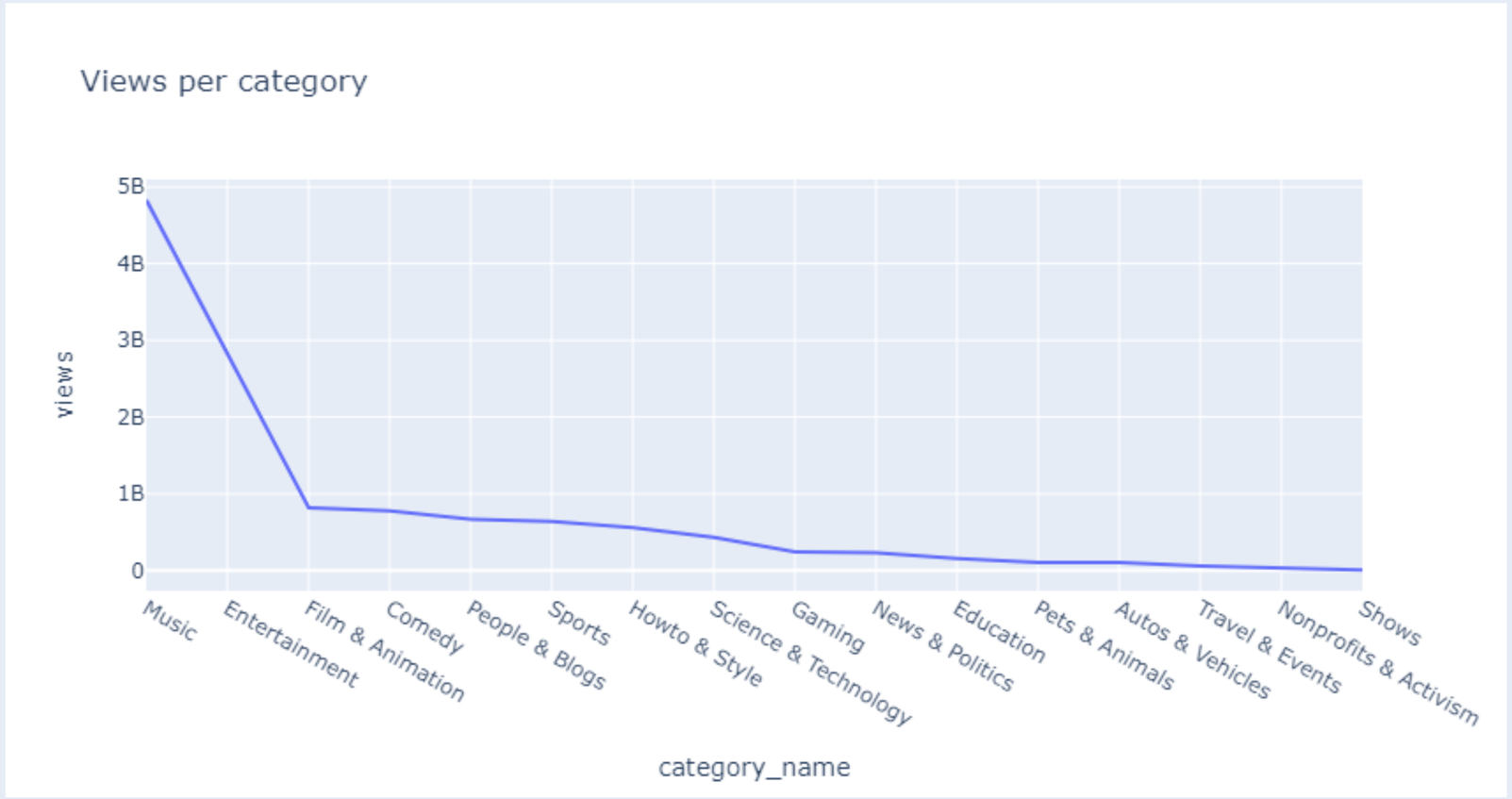
Metodología

- Exploración de los datos
- Limpieza y eliminación de columnas
- Verificación de la coherencia de los datos
- Análisis de datos faltantes y duplicados
- Elección de método para lidiar con NaN values
- Valores en columnas tipo text con caracteres validos para el protocolo ASCII.
- Exportar datasets limpios
- EDA
- Elección y creación de los modelos estadísticos y/o predictivos.

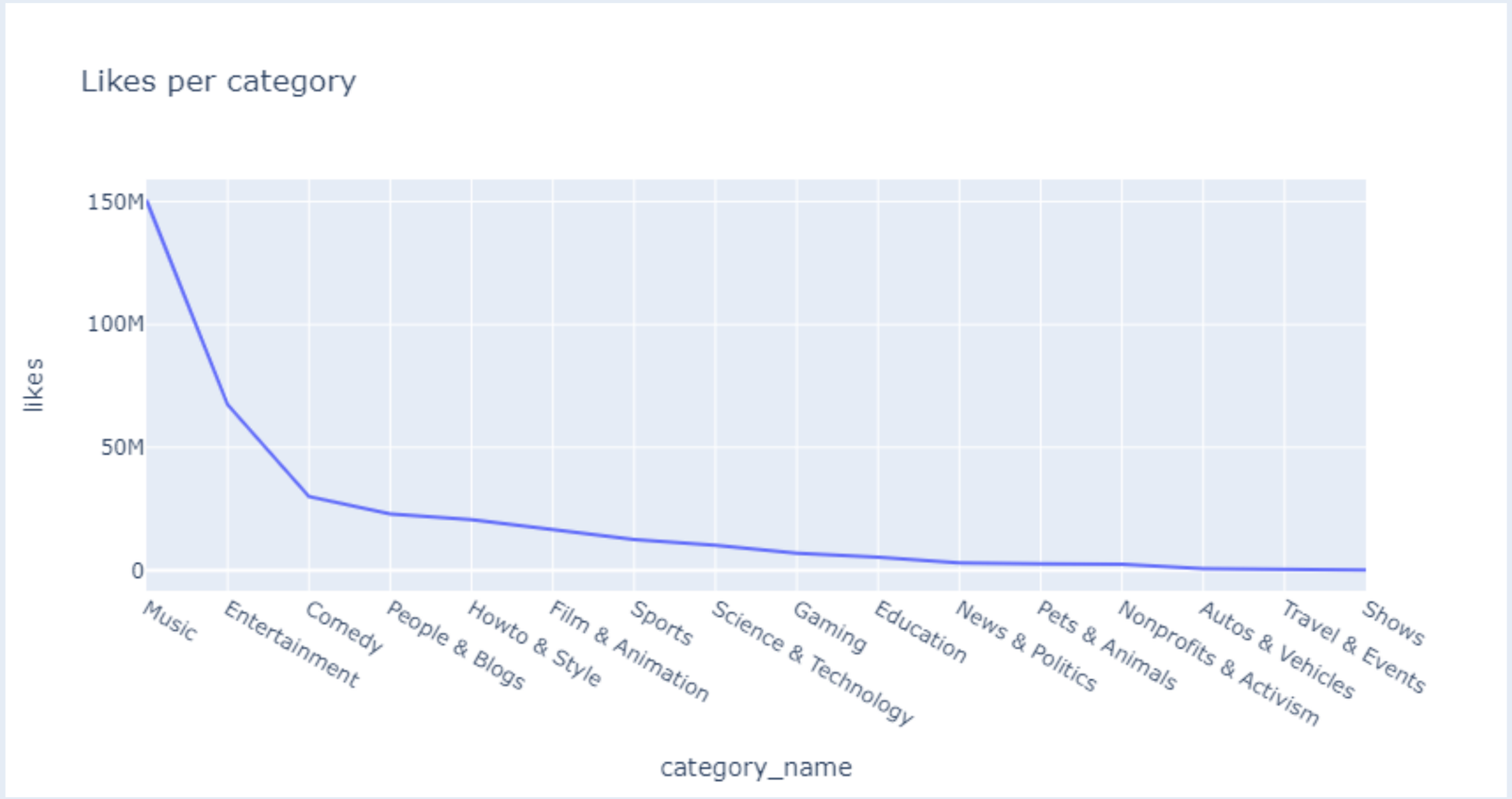
¿Cuáles son las categorías de vídeos que reciben mayores vistas y Likes?

Music es la categoría con mas *views* acumulados

Music es la categoría con mas *likes* acumulados históricamente



- **Music es la categoría que mas vistas acumula con mas de 5 billones de visitas**
- **Entertainment es la categoría que se acerca mas a music pero esta cuenta con un total de 2.8B de visitas acumuladas.**
- **Film & Animation` obtiene el 3er lugar de esta ranking con mas de 814M de visitas acumuladas en sus videos**

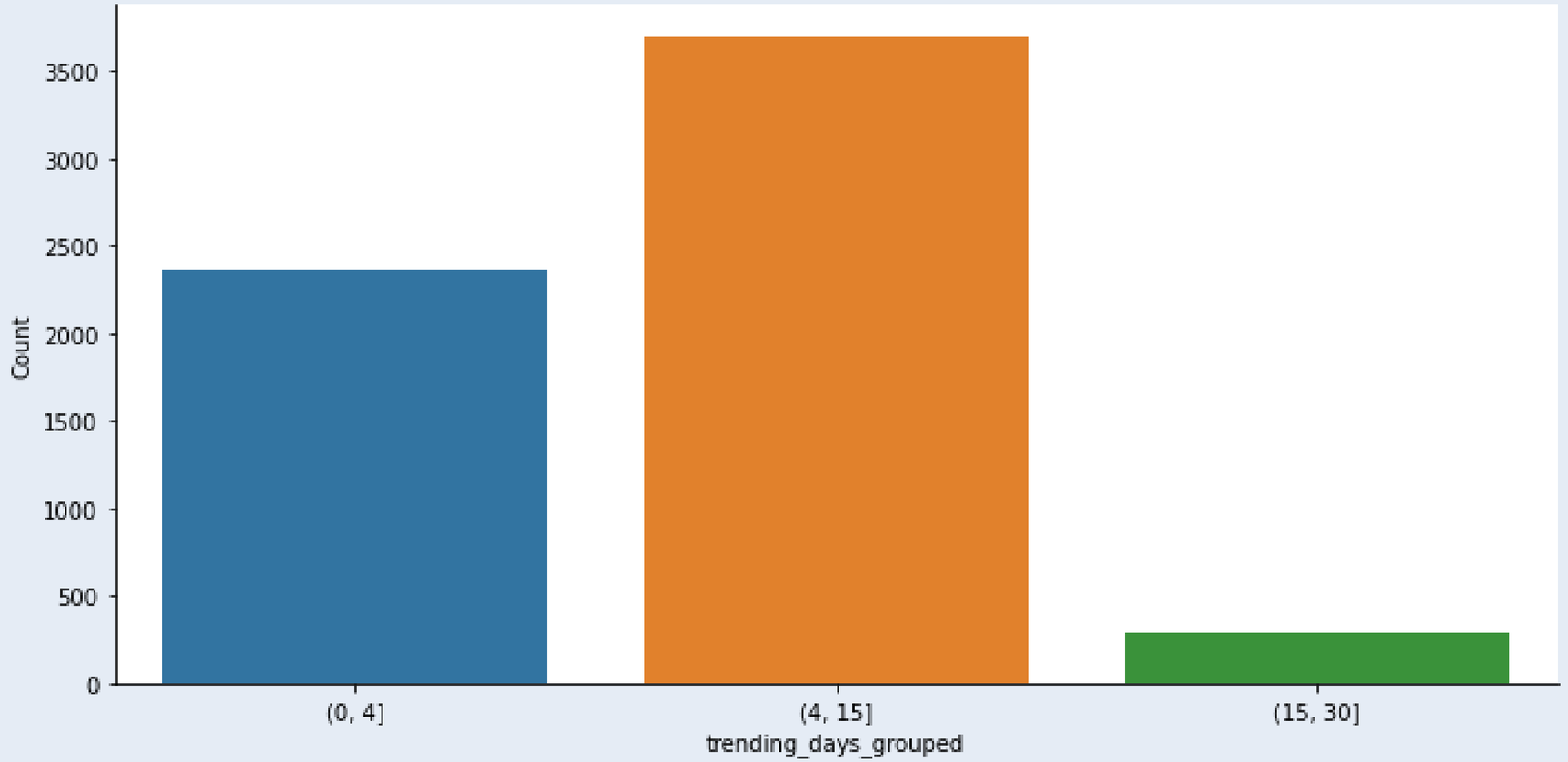


- **Music es la categoría que mas likes acumula con mas de 150 millones de likes**
- **Entertainment es la categoría que se acerca más a music pero esta cuenta con mas de 67 millones de likes acumulados.**
- **Comedy obtiene el 3er lugar de esta ranking con mas de 29 millones de likes acumulados en sus videos.**

¿Es posible encontrar agrupaciones o clasificaciones de videos?

Total de días siendo tendencia

Mejores meses para publicar videos



- **Discretizamos en columnas los valores del total de días que han sido tendencia cada video, agrupando en intervalos de (0, 4], (4, 15] y (15, 30] días, descubriendo que:**
  - **Más de** 3500 **videos consignados en el dataset lograron ser tendencia entre 5 a 15 días**
  - **Como se esperaba es poco común que un video logre ser tendencia de 16 a 30 días, menos de** 500 **videos lo consiguieron.**

¿Cuál es la combinación de características o atributos más importantes que hacen de un video tendencia?

Tags más relevantes por categoría

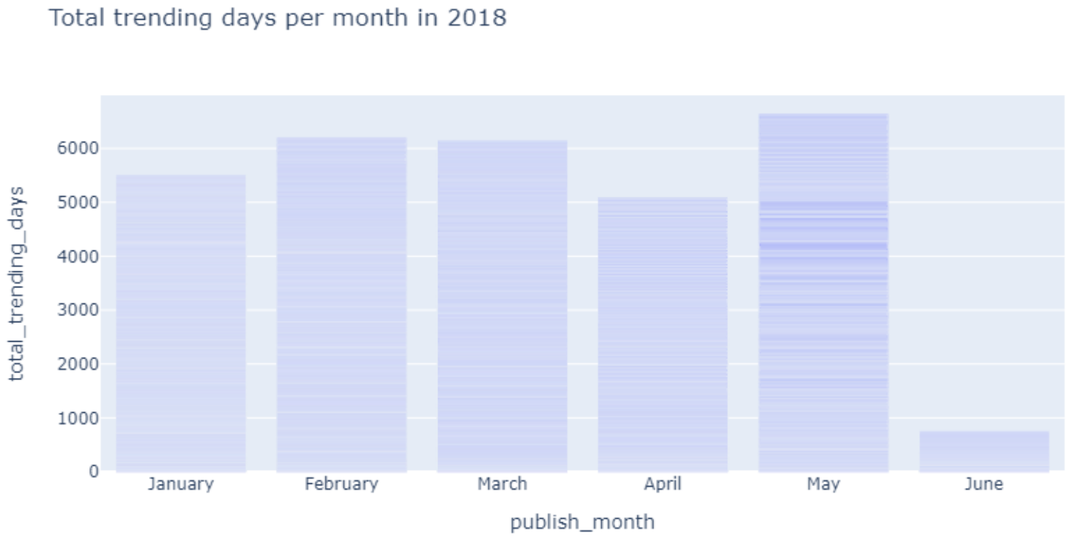
Titulo del video elegido

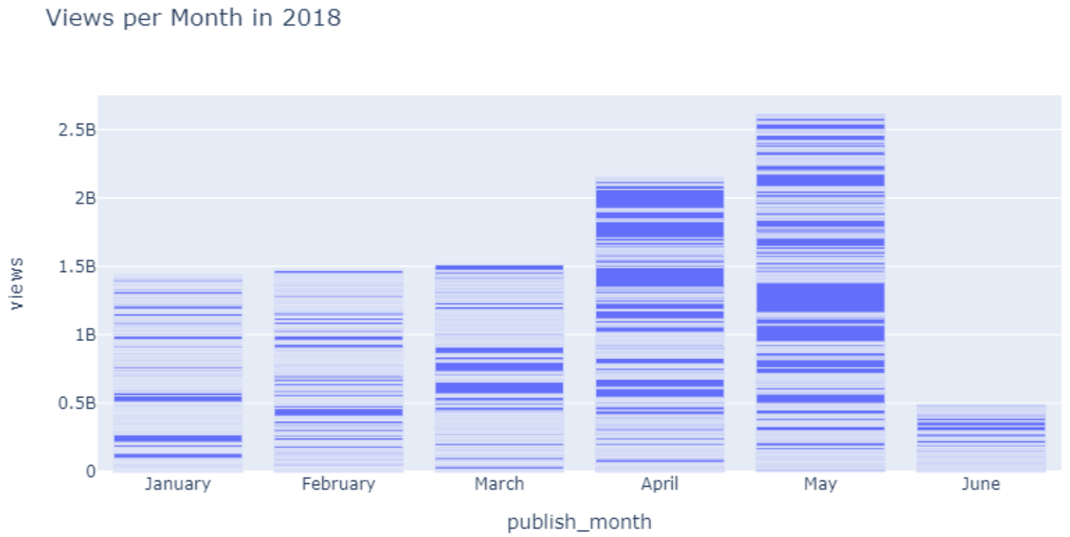
- **Para la categoría elegida (Entertainment) en su histórico de registros, podemos ver palabras como: Marvel (Comics y animación), Jurassic World, Spider-Man, Star Wars, Super Bowl, Sony Pictures, Logan Paul (YouTuber e influencer) son las palabras con mayor frecuencia en videos que se han convertido en tendencia en esta categoría**

¿La temporada o fecha en el que el video es publicado tiene alguna influencia?

Mas que la temporada, conviene analizar las tendencias por categoria

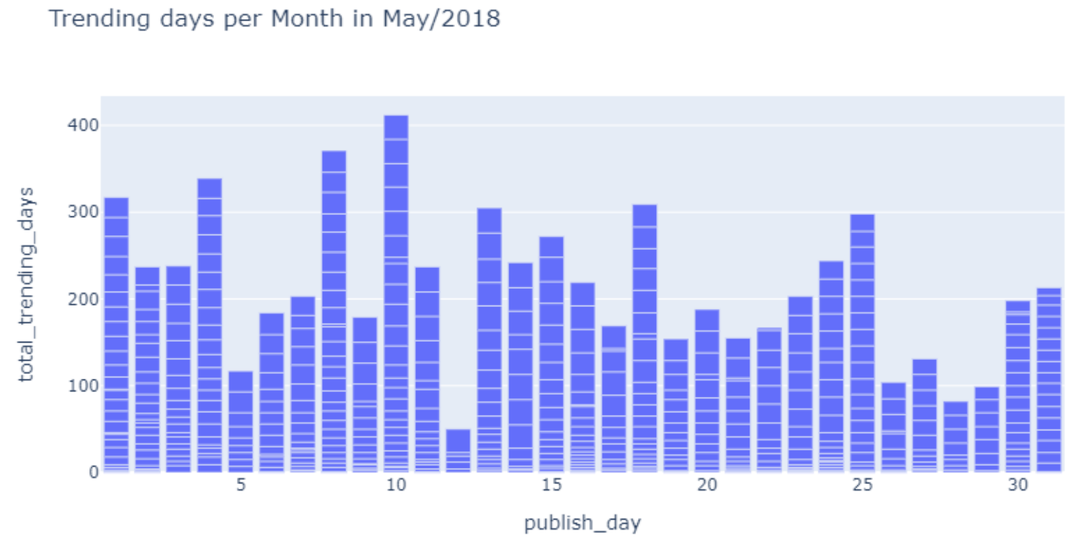
La segmentación del publico objetivo es fundamental

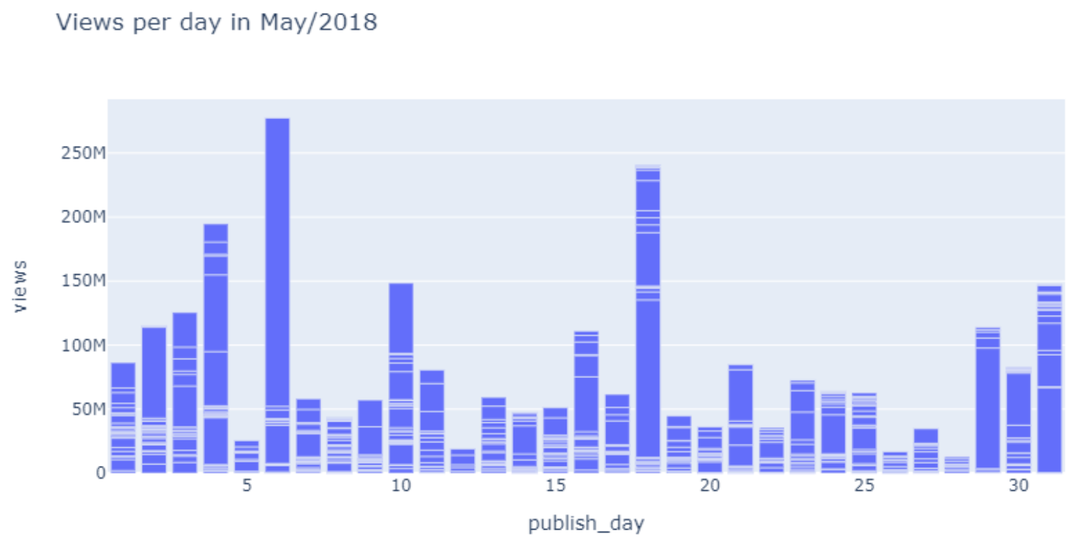




Para el año 2018 *`May`* fue un año idóneo para publicar videos, ya que en este mes se concentro la mayor cantidad de videos que lograron ser tendencias:

- Obtuvieron una mediana de 12 días siendo tendencia
- Obtuvieron una mediana de 1.5 millones de visualizaciones durante el mes





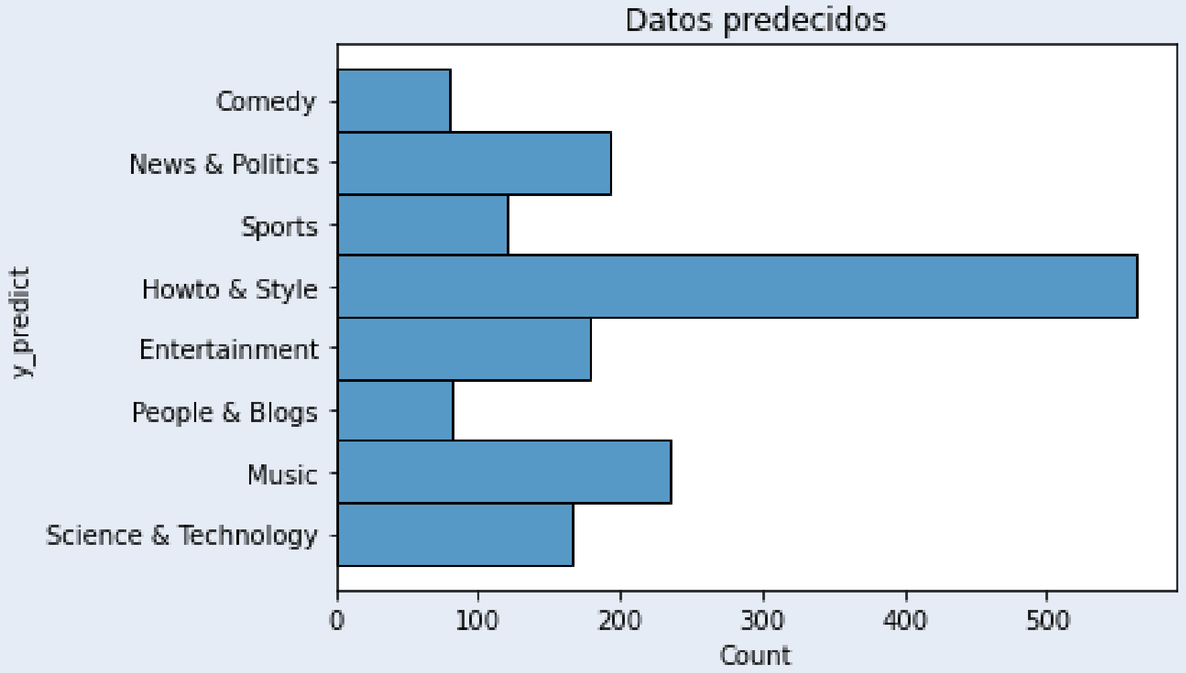
¿Es posible predecir cuantos likes o visitas tendrá un video? Si es así, crea un modelo que lo compruebe.

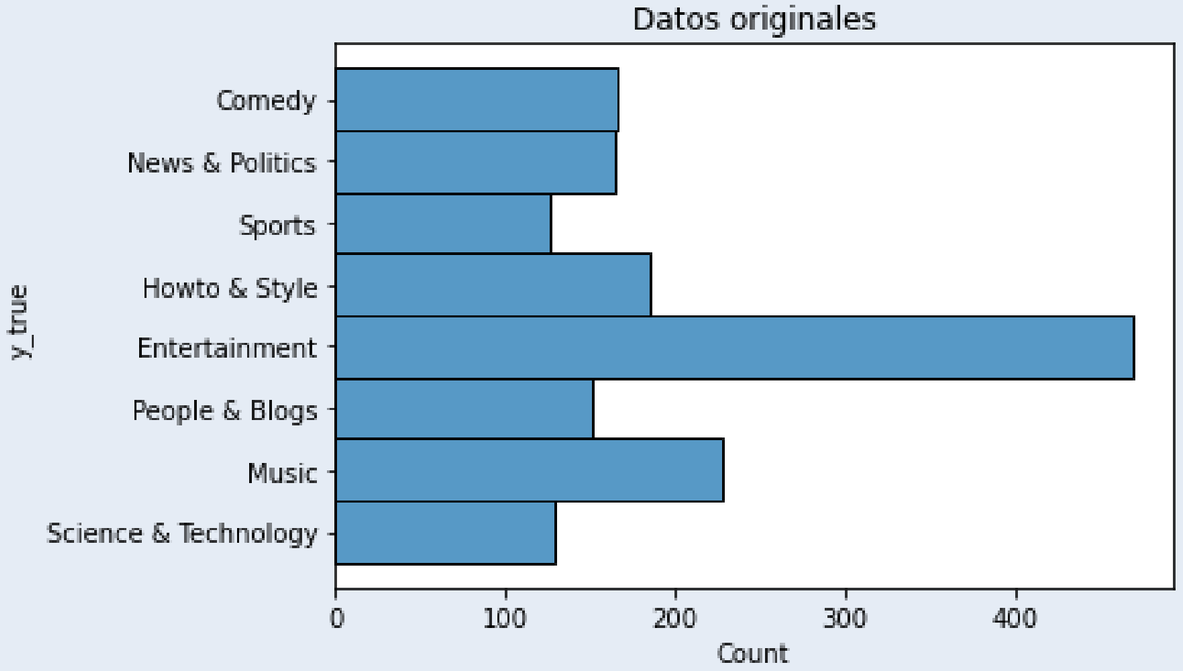
Es posible realizar una predicción con cierto grado de tolerancia como error.

Clasificar el contenido que queremos mostrar, puede ayudarnos a elegir la mejor fecha de publicación

Conclusiones - PREDICCIÓN DE LIKES

- En los primeros modelos se usaron 11 features, correspondientes a las columnas seleccionadas en la etapa de EDA, pero este arrojaba un MAE superior a 120.000
- Se uso la función get_dummies de pandas para realizar un one-hot encoding sobre la columna category_name y de esta manera obtener mejores resultados: el MAE descendió a 60.000 (30.000 en el mejor de los casos)
- Se podría seguir usando mas features por ejemplo: modulo `Polynomical` de sklearn, pero caeríamos en el abismo de la maldición de la dimensionalidad
- Sería relevante usar otro método de análisis como el de `mutual_information` para reconocer que features son mas útiles para el desarrollo de nuestro modelo
- Se deseaba usar la librería de `kerastuner` para mejorar los hiperparametros de la red neuronal base creada, pero mi maquina de pruebas no soporto la cantidad de procesos, usar GPU's como las gratuitas de Google Colab, puede ser una mejora sustancial
- En general los features existentes en el dataset y sus registros, no son información suficiente para predecir con bajo índice de error la cantidad de likes que puede recibir un video
- Para observar los resultado al detalle seguir el link





Conclusiones – CLASIFICACIÓN POR CATEGORÍAS

- El modelo de bolsa de palabras muestra una alta eficiencia para optimizar nuestro modelo clasificador, a pesar de la baja cantidad de registros por categorías
- El estimador `RandomForestClassifier` demuestra una gran adaptación a los datos, usando los parámetros por defecto, implementando regularización, es posible mejorar el score obtenido de 79%
- **Para observar los resultado al detalle seguir el link**