77.50 Introducción a los sistemas inteligentes

Trabajo Práctico Final

Integrantes:

Alumno	padron
Llauró, Manuel Luis	95736
Rial, Sebastián Andrés	90309
Blanco, Sebastian Ezequiel	98539

Fecha de Entrega: 24/06/2019

GitHub:

https://github.com/BlancoSebastianEzequiel/7750-TPFinal

Índice

1. Fase 1: Comprensión del negocio

1.1. Determinar los objetivos del negocio

1.1.1. Escenario actual

Al momento de escribir este informe, la empresa en cuestión tiene interés en desarrollar un canal de youtube, y requiere medir de alguna forma las probabilidades de éxito de llevar a cabo videos populares.

1.1.2. Objetivos del negocio

El objetivo es en este caso poder cuáles son las condiciones que un canal de youtube y sus videos debe cumplir para que los mismos sean exitosos en la plataforma.

1.1.3. Criterios de éxito del negocio

El proyecto se considerará exitoso si se llegan a detectar las variables clave que influyen en que los videos del canal de youtube sean exitosos o no, y en qué nivel influye cada una de ellas. Se considerará que los videos del canal de youtube si los mismos obtienen un porcentaje de X likes y más de 2 millones de vistas por video.

1.2. Evaluación de la situación

1.2.1. Inventario de recursos

Para desarrollar el presente proyecto se cuenta con una amplia gama de recursos que asegura un desarrollo de calidad y confianza del mismo. Se cuenta con un set de datos extraído de la pagina de kaggle sobre videos de youtube que fueron subidos a la plataforma. Además se cuenta con herramientas de software de análisis y visualización de datos líderes en el mercado, como Pandas. Por último, se cuenta con personal altamente calificado para la correcta interpretación de los mismos.

1.2.2. Requisitos, supuestos y restricciones

1.2.2.1 Requisitos

Contar con datos suficientes y sobre todo representativos de la plataforma youtube

1.2.2.2 Supuestos

Los datos en estudio son lo suficientemente correctos como para poder sacar conclusiones confiables a partir de ellos.

1.2.2.3 Restricciones

Se cuenta solamente con datos de videos subidos a youtube Estados Unidos. No se asegura que los resultados sean válidos para otros países.

1.2.3. Riesgos y contingencias

Si bien se puede llevar a cabo un análisis lo más riguroso posible, siempre existe la posibilidad de que un video pueda no ser exitoso porque hay que tener en cuenta que el exito de cada video puede depender de muchos factores de incertidumbre, y si bien se cumplen patrones, no hay nada que asegure al 100En caso de detectar que un video no está teniendo el éxito esperado, habrá que recurrir a las diversas métricas que pueda ofrecer Youtube a desarrolladores y analizar la situación

1.2.4. Terminología

1.2.4.1 Glosario de términos del negocio

Youtube: es una plataforma que te permite subir videos a partir de la creación de un canal.

Canal de youtube: Es el equivalente a crearse una cuenta en cierta aplicacion, en la cual se almacenan los videos subidos

Categoría: define el nombre de un grupo de vides con cualidades comunes.

likes/dislikes: Es un atributo de un video que muestra la cantidad de usuarios que le dieron like al video en cuestion

Vistas: Es la cantidad de veces que el video fue visto

Video exitoso: Aquel cuya cantidad de vistas es mayor a 200000 y promedio de likes versus dislikes es mayor al 80 %

1.2.4.2 Glosario de términos de la minería de datos

Atributo: dato sobre alguna característica de las observaciones.

Atributo relevante: atributo que juega un papel principal en la clasificación,

por lo que la clase dependerá en alguna medida de qué valor tenga.

Registro: fila que representa una observación, está compuesto de atributos.

Dataset: conjunto de datos a ser utilizados para la ejecución de los algoritmos de Data Mining, está compuesto de registros.

Filtrado de atributos: específica al dataset formado considerando sólo los atributos relevantes.

Regla: es una implicación, que representa una acción mediante una condición. Sigue la estructura "Si..., entonces...".

Soporte: es la relación entre la cantidad total de registros del dataset que cumplen la regla y la cantidad de observaciones procesadas.

Confianza: es la relación entre la cantidad total de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones que fueron afectadas por esa misma regla.

Captura: es la relación entre la cantidad de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones procesadas pertenecientes a esa misma clase.

Coeficiente de correlación de Pearson: es la estadística de prueba que mide la relación estadística, o asociación, entre dos variables continuas. Es conocido como el mejor método para medir la asociación entre variables de interés porque se basa en el método de covarianza. Da información sobre la magnitud de la asociación, o correlación, así como la dirección de la relación

1.2.5. Costos y beneficios

El beneficio del proyecto es detectar las características que hacen a un video de youtube sea exitoso. De esta manera se pueden tener en cuenta ciertos parametros para poder desarrollar un video que exitoso basandose en la historia. Al ser un trabajo final educativo, no hay costos.

1.3. Determinar objetivos de Minería de Datos

1.3.1. Objetivo de minería de datos

El objetivo de minería de datos es el análisis de los datos obtenidos a partir de la información disponible, buscando obtener así información relevante que permita predecir las condiciones bajo las cuales una aplicación es exitosa.

1.3.2. Criterios de éxito de minería de datos

Selección de al menos 4 reglas con soporte mayor o igual al 20

1.4. Realizar el Plan del Proyecto

1.4.1. Plan de proyecto

- Recolección de datos: 5 horas.
- Preparación de datos: 5 horas.
- Ejecución del algoritmo de Inducción: 4 horas.
- Análisis de resultados de algoritmo de Inducción: 4 horas.
- Combinación de resultados: 3 horas.
- Elaboración de reporte: 7 horas.

1.4.2. Validación inicial de las herramientas

Se utilizarán las siguientes herramientas:

- Python
- Pandas
- Weka
- Jupiter Notebook
- Numpy

2. Fase 2: Comprensión de los datos

2.1. Recolectar los datos Iniciales

Los datos utilizados durante el transcurso del proyecto fueron obtenidos gratuitamente en el sitio de Kaggle y el mismo se encuentra en formato csv.

2.1.1. Reporte de la recolección de datos iniciales

- El data set se encuentra en la carpeta “‘src/data/“‘ del proyecto.
- Se uso pandas como metodo de recoleccion de datos.
- Para poder levantar el set de datos se usa como separado una coma.

2.2. Descubrir datos

2.2.1. Reporte de descripción de datos

Se cuenta con dos datasets: por un lado se tienen datos estadísticos y generales de los videos descritos en 16 columnas; por otro lado hay un archivo json con el nombre de cada categoria segun el numero asignado. A continuación, los nombres de los features con su correspondiente descripción y tipo de dato:

Mombre de variable	tipo de dato	Descripcion
video_id	alfanumerica	identifica a cada video
trending_date	date	Es una fecha en un dia especifico luego de la publicacion del video
title	alfanumerica	Es el titulo del video
channel_title	alfanumerica	Es el nombre del canal del usuario que publico el videoo
category_id	numeric	it is the number that represent a category nameo
publish_time	timestamp	momento exacto de la publicacion del videoo
tags	alfanumerico	Etiquetas que se pueden agregar al videoo
views	numeric	Cantidad de vistas del video en el momento de la fecha de trending_dateo
likes	numeric	Cantidad de me gusta que el video tiene dados por usuarios de la plataforma youtubeo
dislikes	numeric	Cantidad de no me gusta que el video tiene dados por usuarios de la plataforma youtubeo
comment_count	numeric	Cantidad de comentarios del videoo
thumbnail_link	url	enlace ao
comments_disabled	booleano	Dice si tiene o no los comentarios habilitadoso
ratings_disabled	booleano	Dice si tiene o no los likes/dislikes habilitadoso
video_error_or_removed	booleano	Dice si el video sufrio un error o fue borrado en el momento de la fecha de trending_dateo
description	alfanumeric	Texto escrito por el creador del video que describe el mismoo

2.3. Exploración de los datos

2.3.1. Reporte de exploración de datos

2.3.1.1 Analisis de categorias

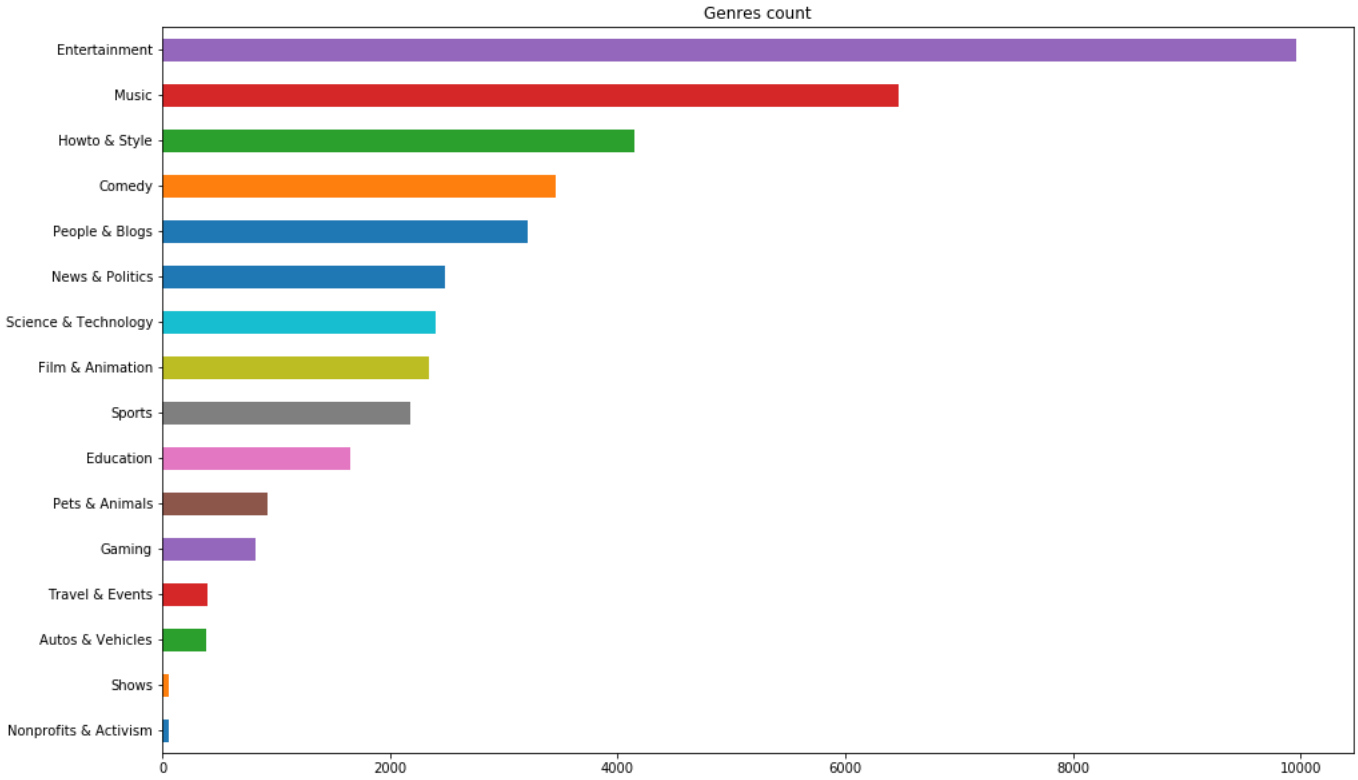

Figura 1: Cantidad de videos segun la categoria

2.3.1.2 Analisis de vistas en funcion de likes

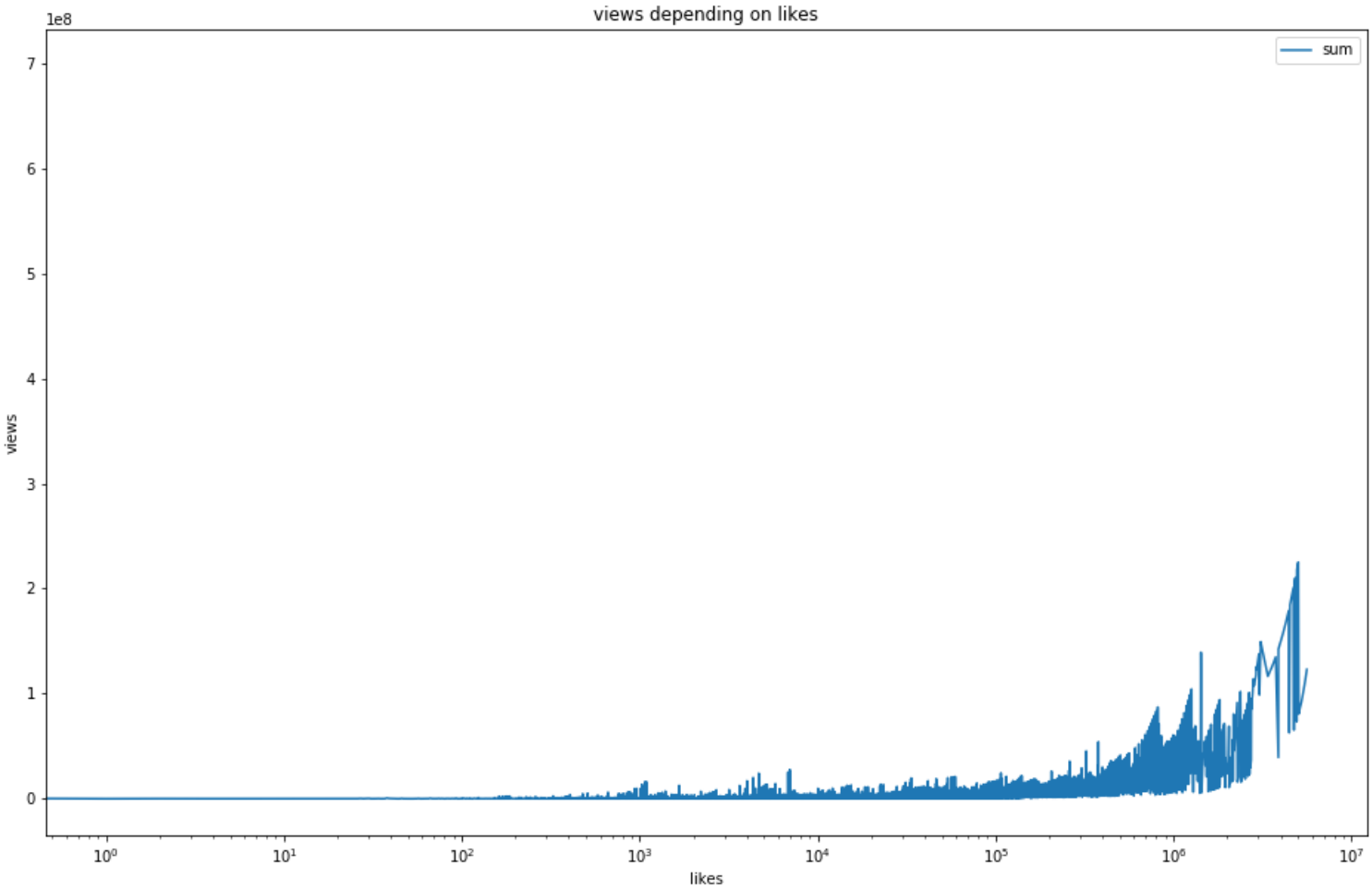

Figura 2: Cantidad de vistas segun la cantidad de likes

Podemos observar que hay un punto de quiebre en el cual a partir de cierta cantidad de likes los videos tienden a aumentar su cantidad de vistas

2.3.1.3 Analisis de vistas en funcion de la diferencia de dias entre la fecha de tendencia y la de publicacion

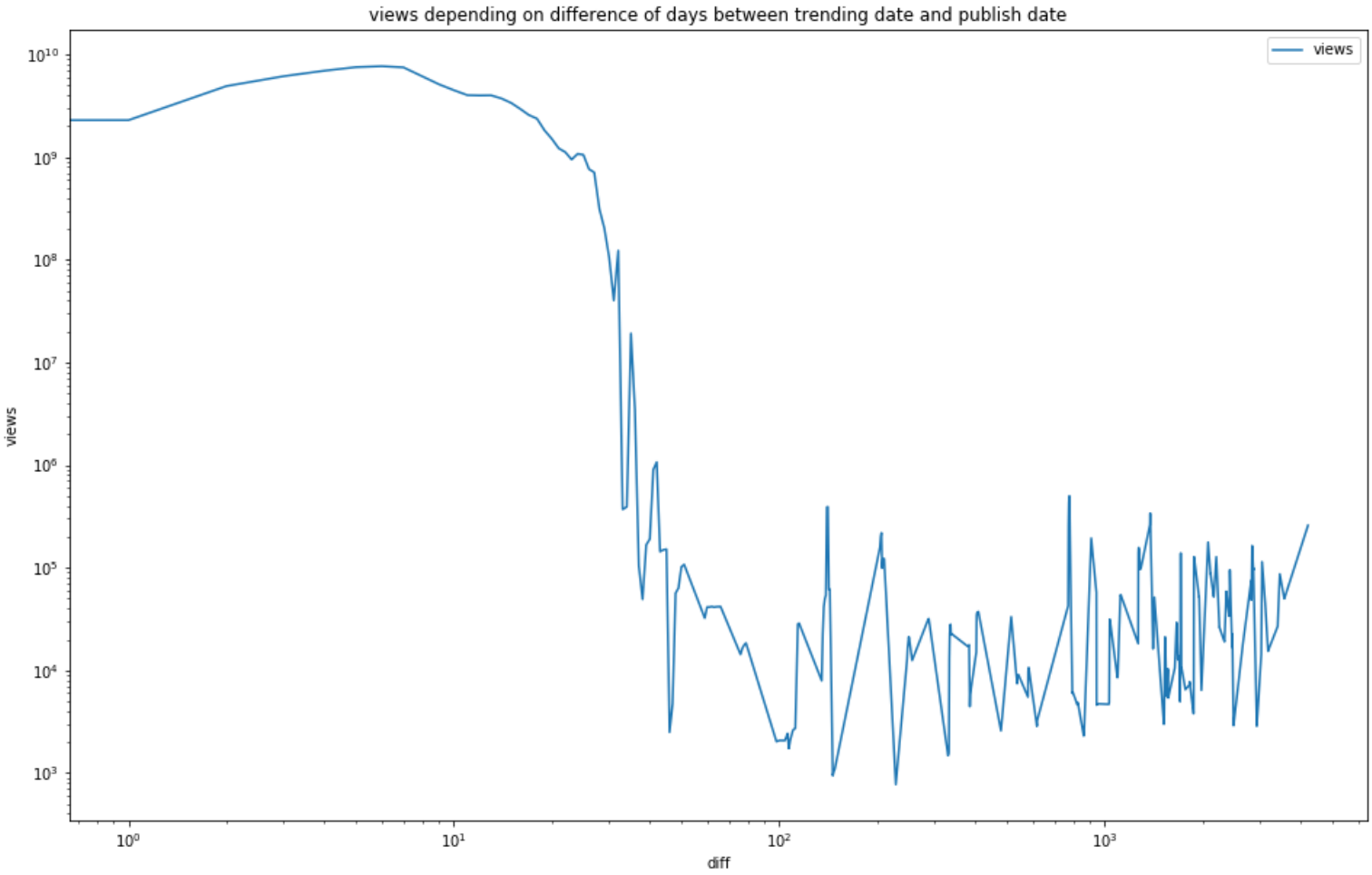

Figura 3: Cantidad de vistas a medida que pasan los dias de su publicacion

Podemos observar que el mayor incremento en la cantidad de vistas es a los pocos dias de la fecha de publicacion del video. Luego vemos que las vistan tienden a decaer

2.3.1.4 Analisis de vistas en funcion de la diferencia de likes y dislikes

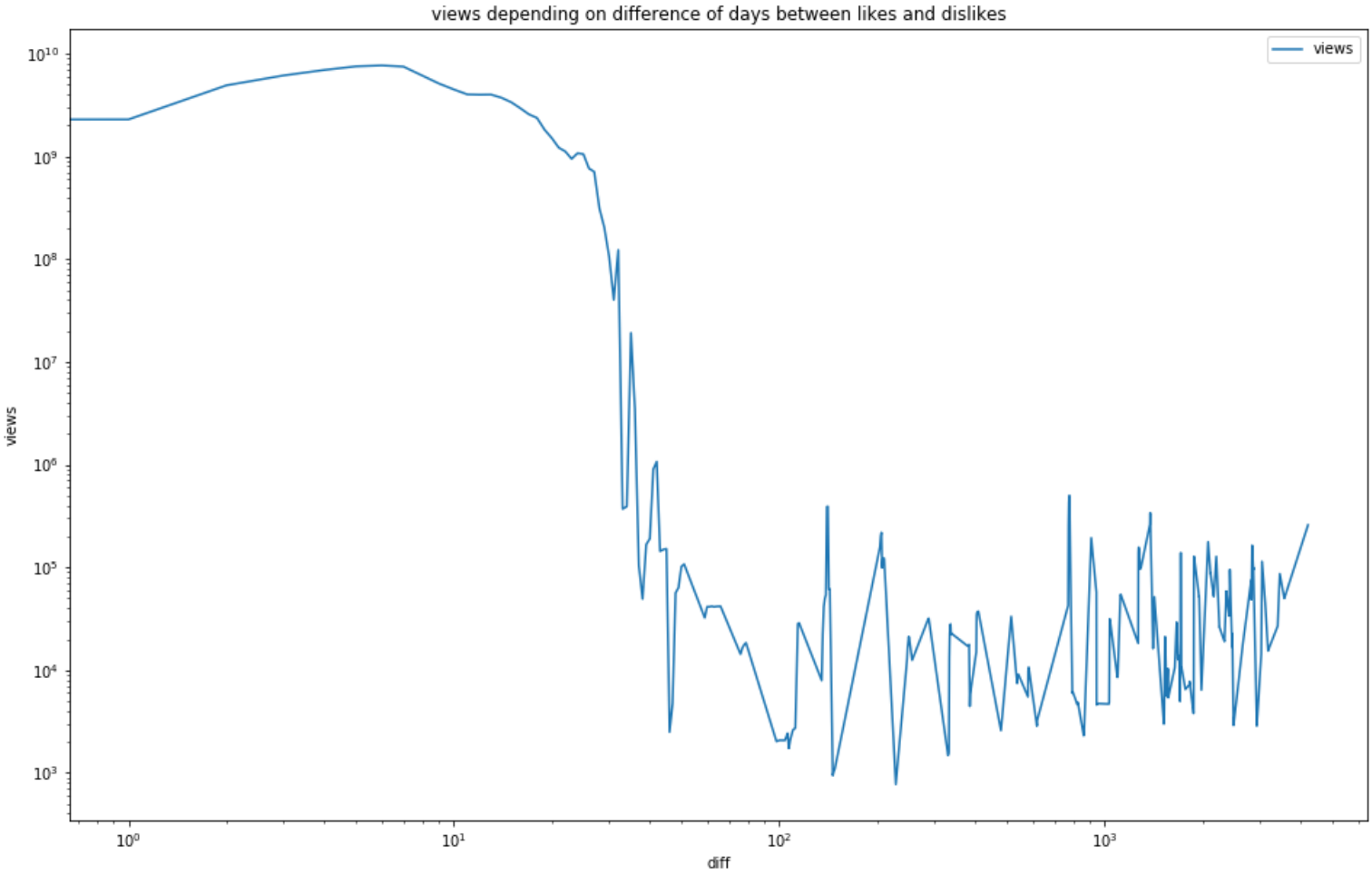

Figura 4: Cantidad de vistas segun la diferencia entre likes y dislikes

Podemos observar que el mayor incremento en la cantidad de vistas es cuando la diferencia entre likes y dislikes no es muy grande. Este grafico puede ser muy capsioso ya que tenemos que tener en cuenta que hay videos de muchas vistas con una diferencia de likes y dislikes grande. Esto quiere decir que esta diferencia es relativa ya que hay que habria que analizar como es una diferencia grande respecto de una pequeña

2.3.1.5 Analisis del progreso de views del video con mayor views del set de datos.

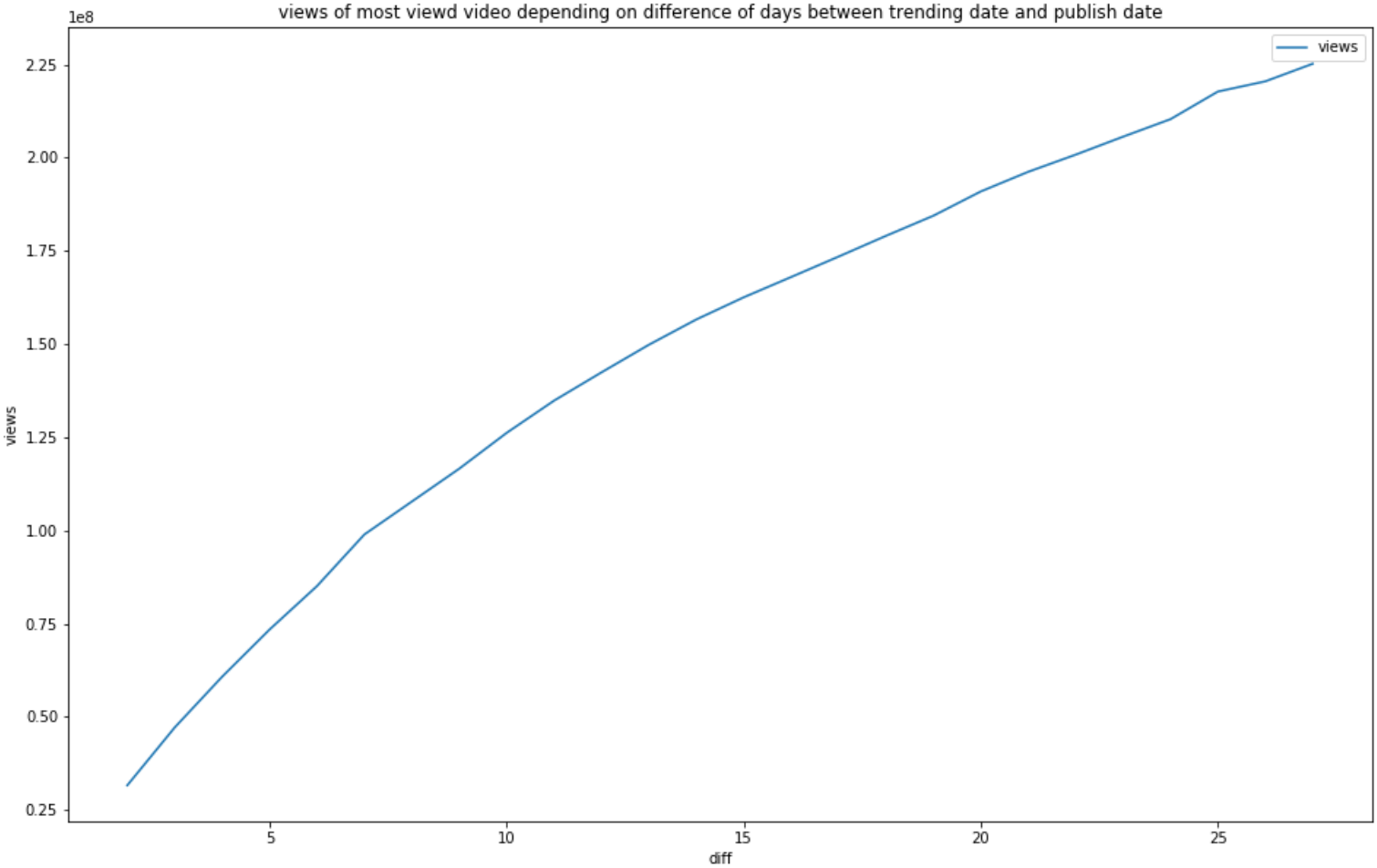

Figura 5: Progreso en dias de las vistas del video mas vistas

Podemos observar que a medida que pasan los dias respecto de la fecha de publicacion del video con mas vistas del set de datos, las vistas del mismo suben linealmente, lo cual tiene sentido ya que cuando un video es exitoso, sus vistan tienen a aumentar progresivamente.

2.3.1.6 Analisis del progreso de views de los videos con mayor, mediano y menor vistas del set de datos.

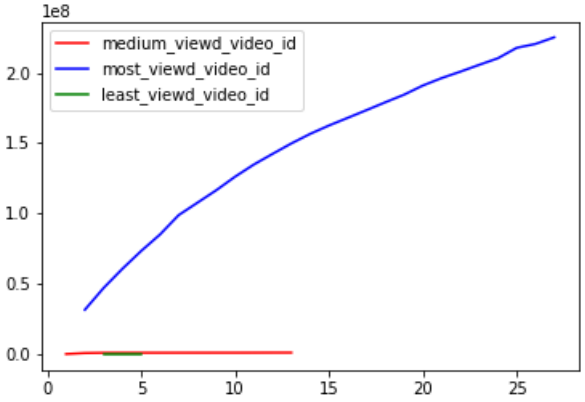

Figura 6: Progreso en dias de las vistas del video mas, mediano y menos vistas

Podemos observar que el video menos visto y el video cuya cantidad de vistas es la mediana del set de datos tienen un progreso de variacion de vistas en el tiempo parecido. Y el video mas visto muestra una amplia diferencia. Esto tiene sentido ya que para tener un progreso como el que tiene el video mas visto se tiene que tener ciertas caracteristicas parecidas. Esto nos dice que hay un grupo pequeño cercano al video mas visto que logra este progreso. y el resto se parece al menos visto

2.3.1.7 Analisis de la cantidad de views segun el largo del titulo del video.

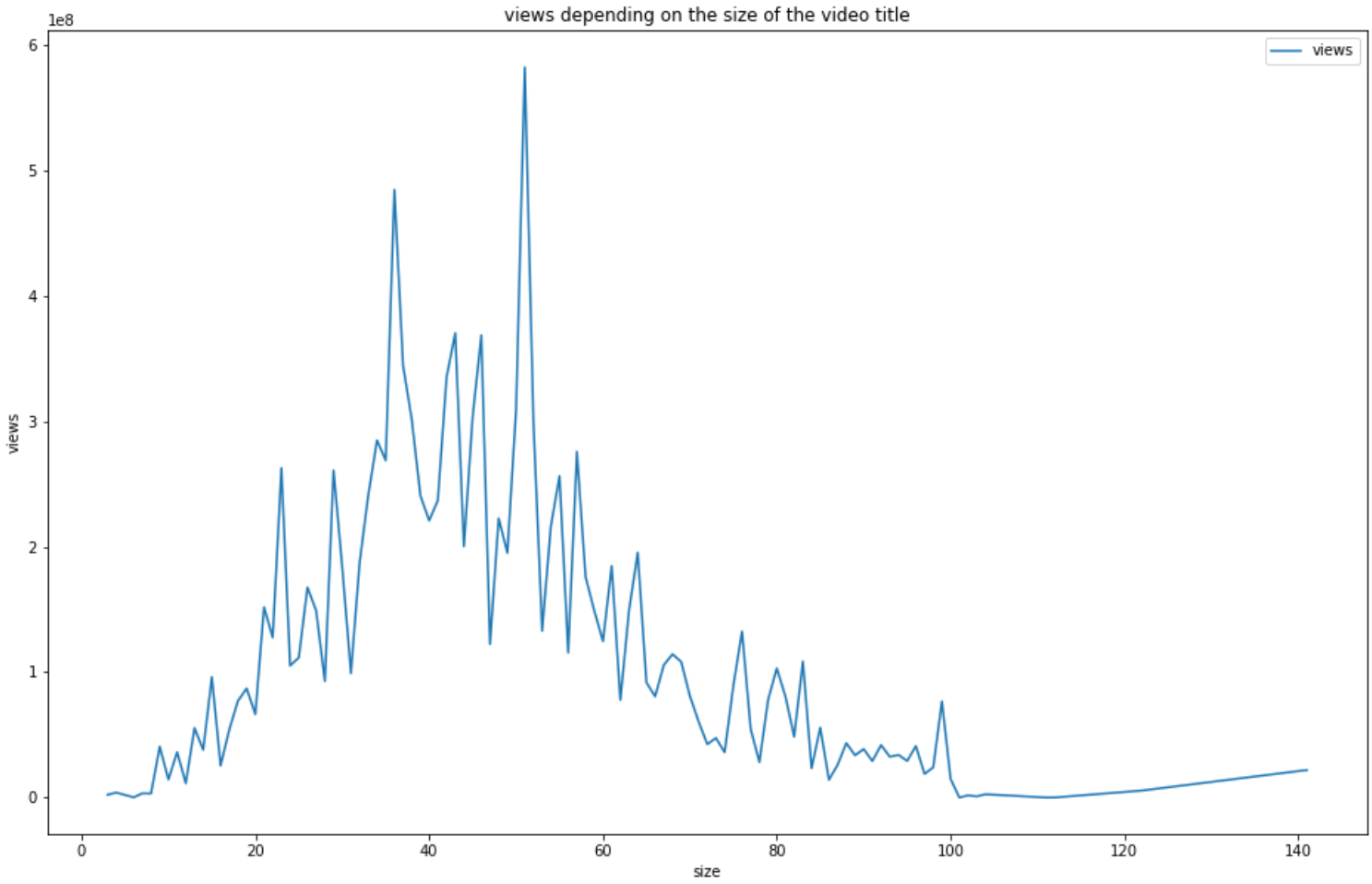

Figura 7: Cantidad de vistas segun el largo del video

Podemos observar que el largo del titulo tiene un rango de valores para su largo en el cual la cantidad de vistas del video logra un valor maximo. Esto tiene sentido ya que el ser humano tiende a leer cosas cortas que generen impacto y es por eso que este rango esta cercano al cero. Tambien si el titulo es muy corto es logico que las vistas no sean altas ya que quizas signifique que ese titulo corto no describe con impacto el video y por lo tanto no genera una atraccion para que la gente lo mire.

2.3.1.8 Analisis la cantidad de videos borrados por categoria.

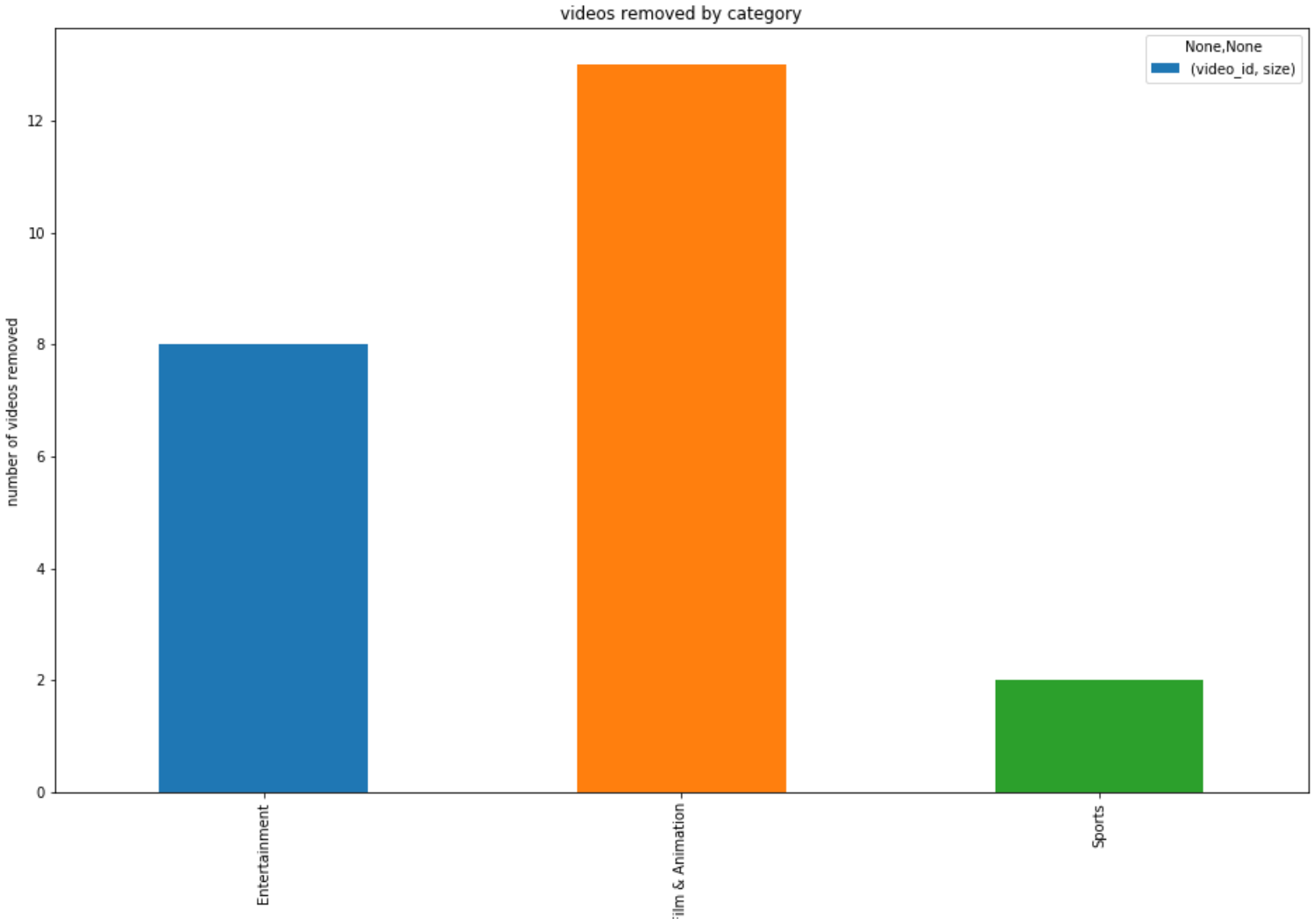

Figura 8: Cantidad de videos borrados por cateogria

Podemos observar que solo en las categorias entretenimeinto, deportes y animacion hay videos que fueron borrados. Esto es probable que se deba a problemas de copywright ya que youtube tiene ciertas resctricciones cuando se muestran imagenes que pertenecen a otras marcas, canciones o peliculas. Youtube ofrece un tiempo maximo para poner una cancion ajena hasta que te bloquean o desmonetizan el video por copywright. Lo mismo ocure con peliculas, o deportes.

2.3.1.9 Analisis de la cantidad de vistas de videos segun tenga o no descripcion.

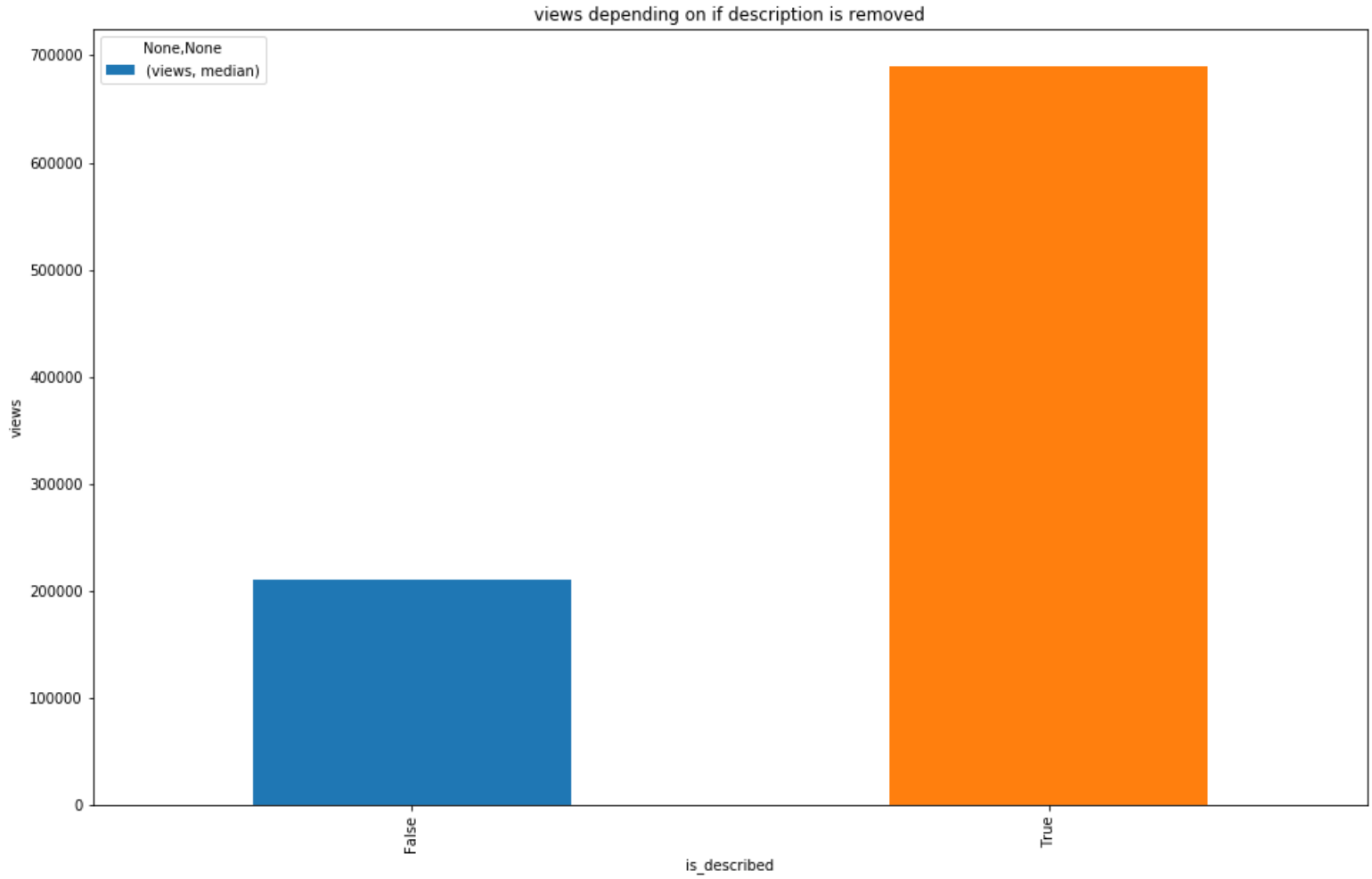

Figura 9: Cantidad de vistas segun su descripcion este o no habilitada

Podemos observar que hay una mayor cantidad de videos con vistas altas en la mediana lo cuales cuentan con una descrpicion. Y los videos que no tiene descripp-cion en general tienden a tener pocas vistas. Esto es probable que se deba a que cuando uno ve una lista de videos en yotube, se tiene una vista previa del mismo, se lee el titulo y un poco la descripcion. Esto es importante a la hora de decidir si uno le interesa ver ese video o no ya que te puede brindar una explicacion que te interese.

2.4. Verificación de calidad de datos

2.4.1. Reporte de calidad de datos

Los datos utilizados durante el transcurso del proyecto fueron obtenidos gratuitamente en el sitio de Kaggle y el mismo se encuentra en formato csv. Con millones de aplicaciones en la actualidad, el conjunto de datos se ha convertido en la clave para obtener las mejores los mejores videos que se publicaron en youtube. Este conjunto de datos contiene más de 40000 detalles de videos publicados en youtube.

Fecha de recolección de datos (de API): Abril 2019

3. Fase 3: Preparación de los datos

3.1. Seleccionar los datos

3.1.1. Inclusión/Exclusión de datos

Los datos que se terminaron utilizando fueron todos los de tipo alfanumerico, numérico y discreto.

3.2. Limpiar los datos

3.2.1. Reporte de limpieza de datos

Se limpiaron los siguientes campos:

- video_id
- trending_date
- title
- channel_title
- tags
- likes
- dislikes
- thumbnail_link
- description

Algunos de estos campos se usaron para crear nuevos campos.

3.3. Estructurar los datos

3.3.1. Derivación de atributos

- **days_since_publication:** Cantidad de dias desde que se publico el video
- **title_size:** Longitud en cantidad de caracteres del atributo title
- **publish_time:** Se modifico el timestamp por un formato tal: YYYY-MM-DD
- **tags_quantity:** cantidad de tags
- **likes_ratio:** Ratio de likes y dislikes calculado como: likes/(likes+dislikes)
- **has_description:** Es verdadero si el video tiene descripcion.

3.3.2. Generación de registros

Se genero el registro `is_sucessfull` que fue calculado en funcion de si el video tiene mas de 2 millones de viws y un ratio mayor o igual a 0.8

3.4. Integrar los datos

3.4.1. Unificación de los datos

Se unieron las datos del dataset con el json de categorias para conocer la relacion entre el valor nuemrico y el nombre de la categoria.

3.5. Formato de los datos

3.5.1. Reporte de formato de los Datos

publish_time: Se modifico el timestamp por un formato tal: YYYY-MM-DD

4. Fase 4: Modelado

4.1. Seleccionar una técnica de modelado

4.1.1. Técnica de modelado

4.1.2. Supuestos de modelado

4.2. Generar el diseño de las pruebas

4.2.1. Diseño de las pruebas

4.3. Construir el modelo

4.3.1. Configuración de parámetros

4.3.2. Modelos

4.3.3. Descripción del modelo

4.4. Evaluar el modelo

4.4.1. Evaluación del modelo

4.4.2. Revisión de la configuración de parámetros

5. Fase 5: Evaluación

5.1. Evaluar Resultado

5.1.1. Valoración de los resultados de minería de datos

5.1.2. Modelo aprobado

5.2. Proceso de revisión

5.2.1. Revisión del proceso

5.3. Determinar Próximos pasos

5.3.1. Listado de posibles acciones

6. Fase 6: Implementación

6.1. Plan de Implementación

6.1.1. Plan de Implementación

6.2. Plan de monitoreo y mantenimiento

6.2.1. Plan de monitoreo y mantenimiento

6.3. Informe Final

6.3.1. Informe Final

6.3.2. Presentación final

6.4. Revisión del proyecto

6.4.1. Documentación de la experiencia realizada

7. Conclusiones