

# Laporan Proyek MATH 1042 – Peluang dan Statistika

## Proyek 1A: Lautan Video Youtube

### I. PENGANTAR

#### A. Latar Belakang

Youtube merupakan salah satu platform paling ramai di dunia maya saat ini. Lebih dari 800 juta video telah beredar di Youtube. Dari banyaknya video tersebut, ada beberapa video yang menerima perhatian global, atau 'viral'/'trending'. Kami akan mengeksplorasi sebuah dataset <https://drive.google.com/file/d/1k39YhjiEye25a1pZbP4H1BzyLKie9QNA/view?usp=sharing> yang berisi 40949 video Youtube yang beredar di Amerika Serikat dan melihat seberapa 'terpencil'nya video-video yang viral tersebut dengan meninjau 3 parameter video berupa:

- Views : Banyaknya pengguna yang pernah menonton video tersebut,
- Likes : Banyaknya pengguna yang menyukai video tersebut,
- Dislikes : Banyaknya pengguna yang tidak menyukai video tersebut.




#### B. Tujuan

- Membuat rangkuman dan visualisasi data video Youtube yang praktis dan nyaman dilihat.
- Melihat seberapa terpencilnya video viral dibanding video biasa.
- Menemukan informasi menarik dari data video-video Youtube.

## II. RANGKUMAN HASIL TUGAS PEMROGRAMAN DAN FAKTA MENARIK




#### A. Hasil Akhir Ukuran Pusat Data dan Ukuran Variasi

## PEMROGRAMAN: UKURAN PUSAT DATA DAN VARIASI

	MEAN	MEDIAN	MODUS	STANDAR DEVIASI
	2.360.785	681.861	[0, 22.521.686]	7.394.114
	74.266	18.091	[0, 561.382]	228.885
	3711	631	[0, 167.442]	29.029

### B. Hasil Akhir Ukuran Lokasi

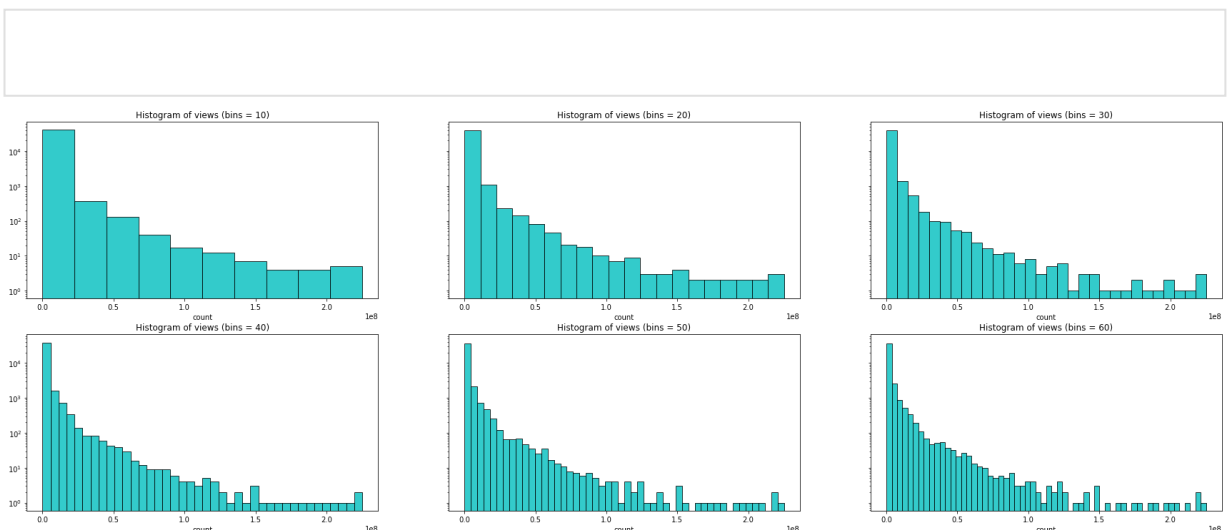
## PEMROGRAMAN: UKURAN LOKASI

	KUARTIL 2	KUARTIL 3	IQR	BATAS BAWAH	BATAS ATAS	JUMLAH PENCILAN
	242.329	1.823.157	1.580.828	-2.128.913	4.194.399	4499
	5.242	55.417	49.993	-69.565,5	130.406,5	5136
	202	1.938	1.736	-2.402	4.542	5288

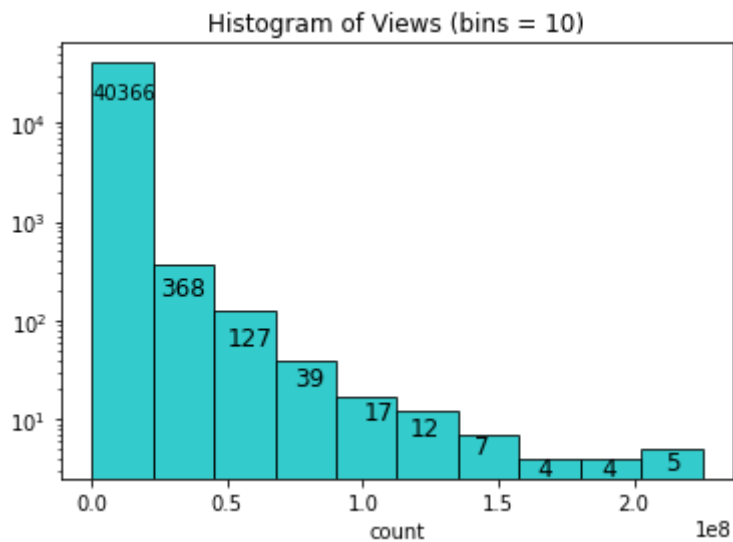
### C. Histogram

#### C.1 Views

In [52]:

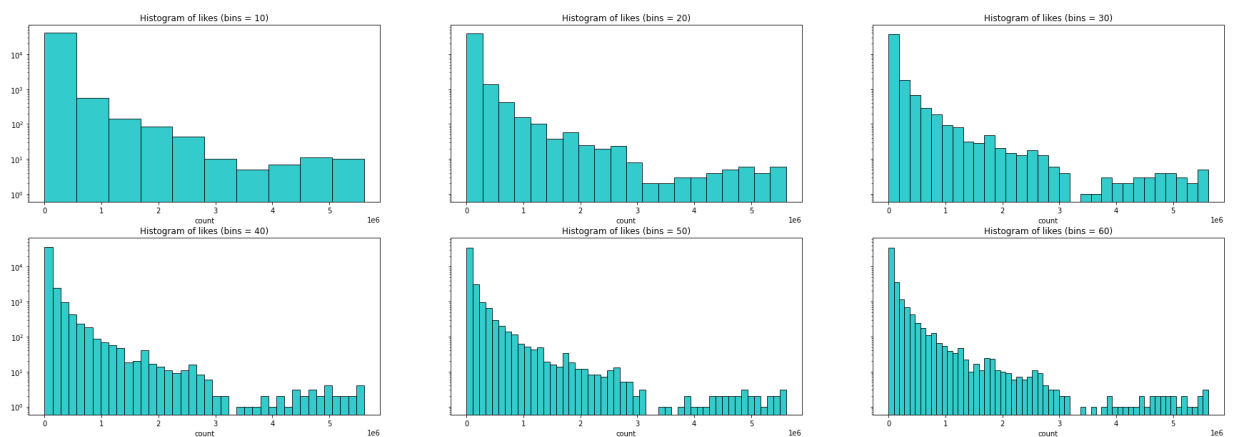


In [53]:

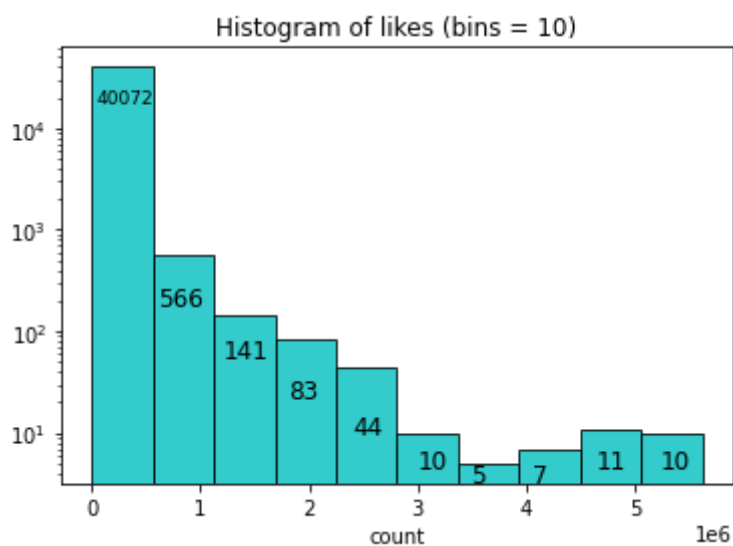


## C.2 Likes

In [54]:

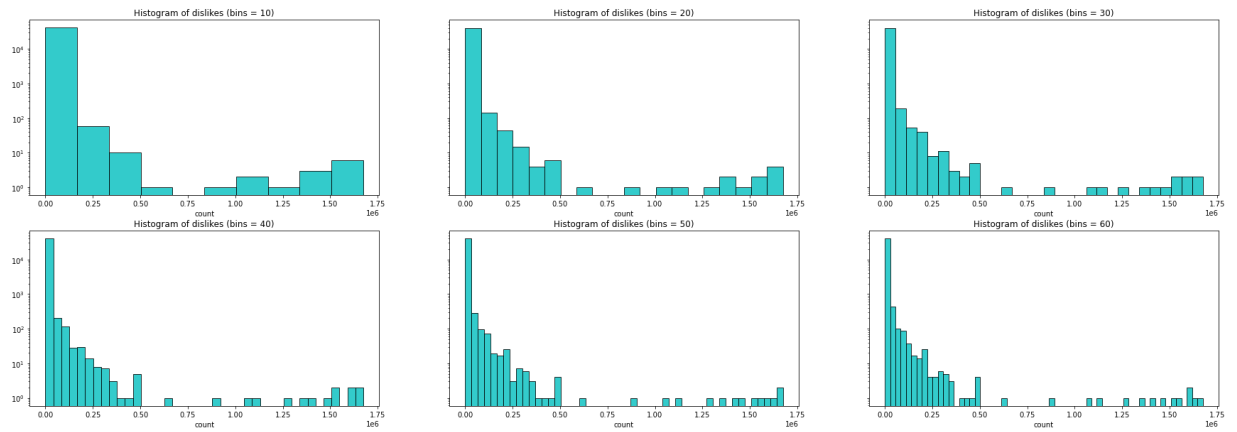


In [55]:

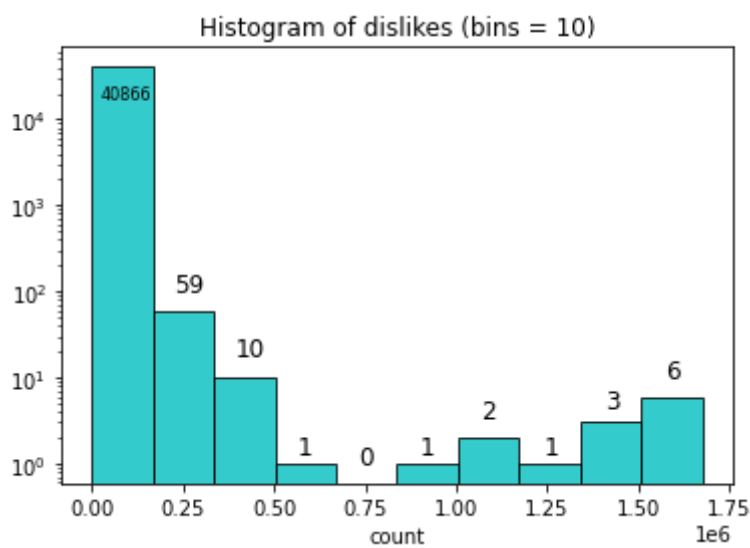


## C.3 Dislikes

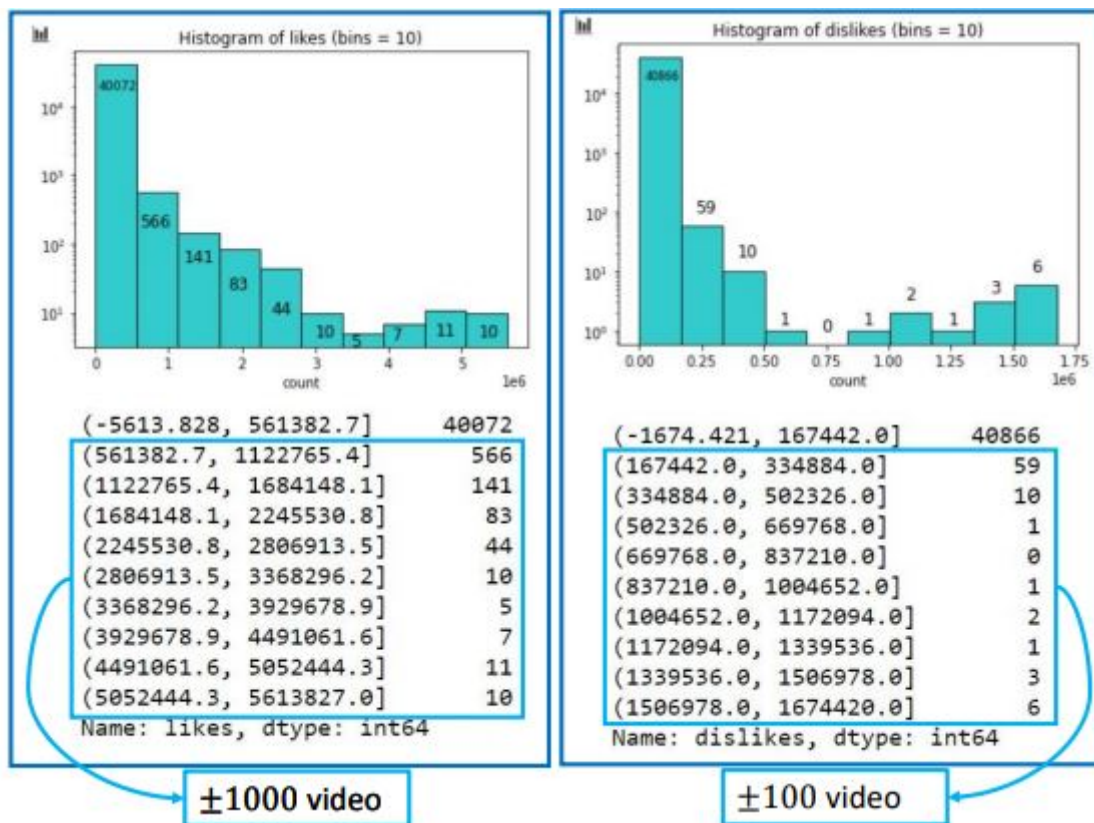
In [56]:



In [57]:



Seluruh histogram dibuat dengan sumbu y berskala logaritmik agar memperjelas visualisasi data-data yang nilainya terlalu kecil. Untuk sumbu x, kami membagi interval sama rata berdasarkan nilai minimum dan maksimum masing-masing parameter. Awalnya kami ingin melihat bagaimana distribusi tiap parameter dengan memanipulasi jumlah bins mulai dari 10 hingga 60. Namun, tetap terlihat bahwa jumlah views, likes, dan dislikes yang paling sedikit tetap paling banyak. Kami memutuskan untuk menggunakan histogram 10 bins karena lebih nyaman dilihat dan merepresentasikan kesimpulan yang sama dengan menggunakan 60 bins. Dari histogram yang ada, dapat disimpulkan bahwa video di Youtube didominasi dengan video-video yang memiliki views, likes, dan dislikes yang di bawah rata-rata.



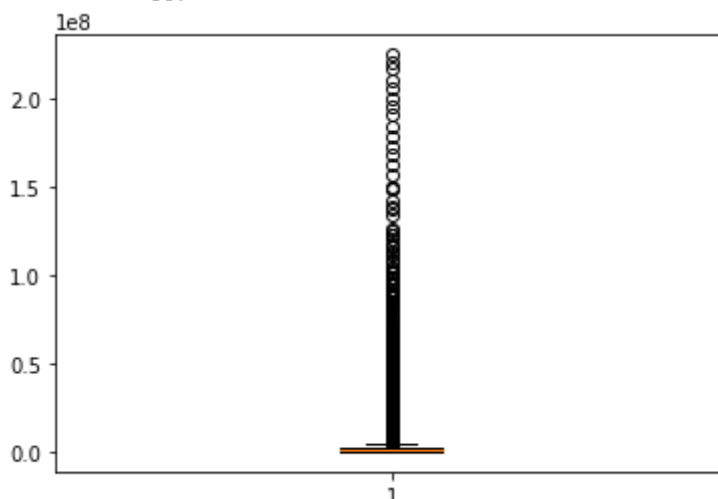
Berdasarkan data histogram likes dan dislikes, kita bisa melihat bahwa video Youtube cenderung memperoleh likes daripada dislikes. Hal ini menunjukkan bahwa Algoritma Filter Bubble youtube bekerja dengan baik. Algoritma Filter Bubble adalah algoritma yang digunakan oleh Youtube untuk memunculkan video yang cenderung lebih disukai oleh penonton.

## D. Boxplot

### D.1 Views

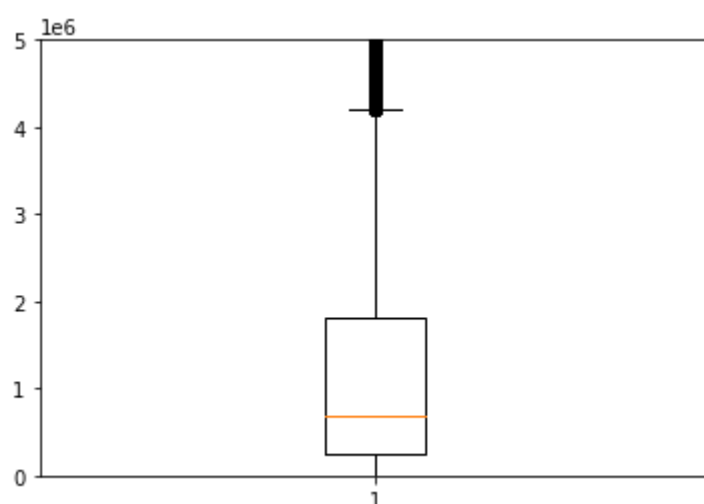
In [58]:

```
Out[58]: {'whiskers': [<matplotlib.lines.Line2D at 0x14d27dcac40>,
<matplotlib.lines.Line2D at 0x14d27dca580>],
'caps': [<matplotlib.lines.Line2D at 0x14d27dca8e0>,
<matplotlib.lines.Line2D at 0x14d2805fdc0>],
'boxes': [<matplotlib.lines.Line2D at 0x14d27dcabe0>],
'medians': [<matplotlib.lines.Line2D at 0x14d2805f4f0>],
'fliers': [<matplotlib.lines.Line2D at 0x14d2805fa90>],
'means': []}
```



In [59]:

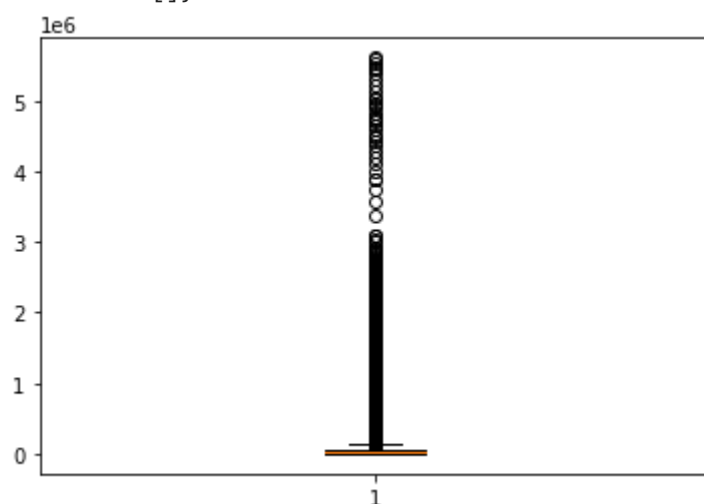
Out[59]: (0.0, 5000000.0)



## D.2 Likes

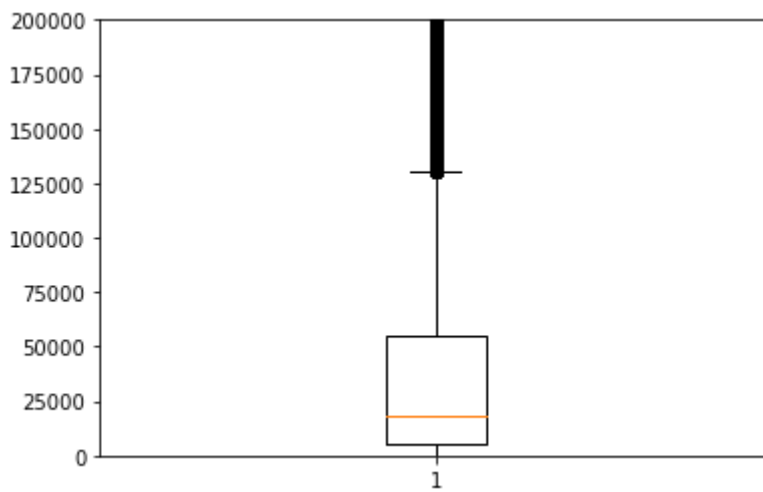
In [60]:

Out[60]: {'whiskers': [<matplotlib.lines.Line2D at 0x14d27ed6250>, <matplotlib.lines.Line2D at 0x14d27ed6cd0>], 'caps': [<matplotlib.lines.Line2D at 0x14d281614f0>, <matplotlib.lines.Line2D at 0x14d28161130>], 'boxes': [<matplotlib.lines.Line2D at 0x14d27ed6880>], 'medians': [<matplotlib.lines.Line2D at 0x14d28161b80>], 'fliers': [<matplotlib.lines.Line2D at 0x14d28161d00>], 'means': []}



In [61]:

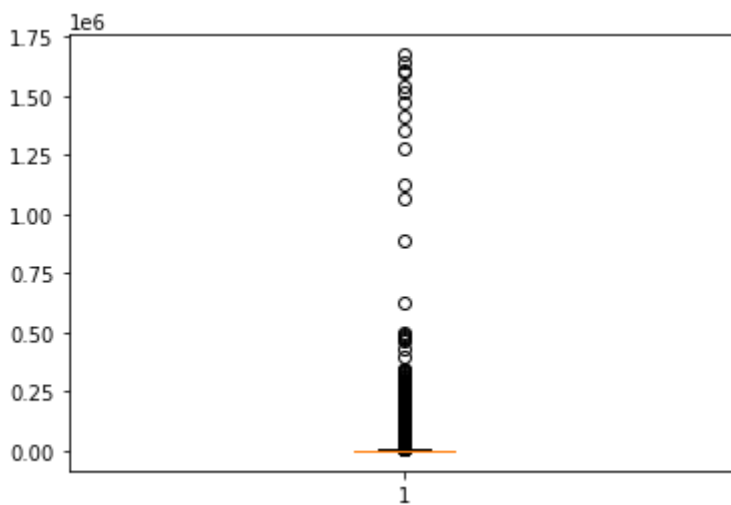
Out[61]: (0.0, 200000.0)



### D.3 Dislikes

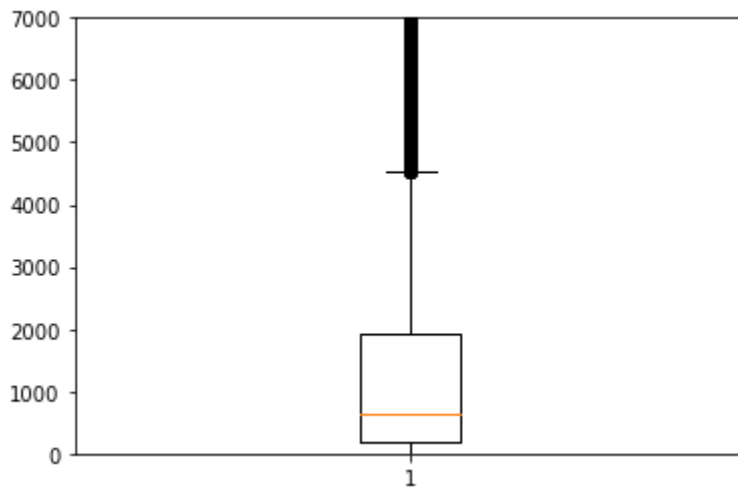
In [62]:

```
Out[62]: {'whiskers': [<matplotlib.lines.Line2D at 0x14d2819a310>,
<matplotlib.lines.Line2D at 0x14d2819abe0>],
'caps': [<matplotlib.lines.Line2D at 0x14d2819ae20>,
<matplotlib.lines.Line2D at 0x14d2819aa00>],
'boxes': [<matplotlib.lines.Line2D at 0x14d28bb0910>],
'medians': [<matplotlib.lines.Line2D at 0x14d2819a250>],
'fliers': [<matplotlib.lines.Line2D at 0x14d28b686d0>],
'means': []}
```



In [63]:

```
Out[63]: (0.0, 7000.0)
```



Dari boxplot, terlihat bahwa baik jumlah views, likes, dan dislikes berpusat di nilai yang rendah. Pencilan yang ada juga sangat banyak dan bernilai sangat jauh dari pusat data. Meskipun jumlah pencilan tidak sebanyak jumlah data di pusat, namun pencilan ini tidak boleh dihilangkan karena nilainya yang besar dapat mengubah seluruh hasil pengolahan data. Pencilan-pencilan di sini adalah video-video yang viral tersebut.

## IV. PERTANYAAN DISKUSI

**a) Seberapa ‘terpencil’/‘jauh di atas sana’ video-video dengan total #views lebih dari 100 juta, dibandingkan dengan data video-video lainnya? Bagaimana Anda mengungkapkan hal ini secara kuantitatif?**

Kami menyatakan keterpencilan video dengan likes lebih dari 100 juta dalam bentuk persentase untuk melihat seberapa sedikitnya golongan ini dibandingkan golongan video lainnya. Caranya mudah, yaitu dengan hanya membandingkan jumlah video dengan likes lebih dari 100 juta dengan jumlah seluruh video. Dari perhitungan didapatkan bahwa dari 40.949 video, hanya 40 yang memiliki views lebih dari 100 juta, atau dapat dikatakan hanya 0.098%.

```
In [64]: more_than_100M_views = len(X[X["views"] >= 100_000_000])
         more_than_100M_views
```

```
Out[64]: 40
```

```
In [65]: all_videos = len(X)
         all_videos
```

```
Out[65]: 40949
```

```
In [66]: percentage_100M_views = 100 * more_than_100M_views/all_videos
         print("Persentase video dengan views lebih dari 100 JT = %.3f" %(percentage_100M_views))
```

```
Persentase video dengan views lebih dari 100 JT = 0.098
```

**b) Anda hendak mengumpulkan video-video dengan jumlah #likes terbanyak dan melabeli kumpulan video tersebut ‘video terfavorit’. Berapakah batas minimal #likes yang perlu Anda tetapkan jika Anda mengkehendaki bahwa hanya terdapat sekitar 0,1% saja video yang dapat tergolong sebagai ‘video favorit’?**

Dalam bagian ini, kami menggunakan fungsi `get_minimum_point` dengan parameter `df`, `key`, dan



proportion. "df" adalah dataframe yang digunakan. Dalam proyek ini, dataframe yang digunakan adalah dataframe X. "key" adalah bagian dari dataframe yang ingin ditinjau nilainya untuk suatu proporsi tertentu. Dalam mencari video terfavorit, "key"-nya adalah parameter likes. "proportion" adalah seberapa banyak data dalam "key" yang ingin dikhususkan. Dalam kasus ini, kekhususannya adalah digolongkan favorit dengan jumlah data sebanyak 0.1% dari seluruh data.

Algoritma get\_minimum\_point akan mulai dengan mengurutkan data, kemudian mencari data pada indeks yang sudah ditentukan berdasarkan proposi yang diinginkan. Jika kita menghendaki bahwa hanya terdapat hanya sekitar 0.1% saja video yang dapat tergolong sebagai video favorit, maka jumlah minimum likes yang harus dimiliki video itu adalah 2,906,264.

```
In [67]: def get_minimum_point(df, key, proportion):
          x = df[key].sort_values(ascending=False)
          return x.iloc[int(len(x) * proportion)-1]
```

```
In [68]: get_minimum_point(X, 'likes', 0.1/100)
```

```
Out[68]: 2906264
```

**c) Video dengan jumlah #likes yang cukup banyak tidak menjadi jaminan bahwa video tersebut berkualitas baik. Sebagai contoh, sebuah video mungkin memiliki jumlah #likes cukup besar, namun rupanya juga memiliki jumlah #dislikes yang lebih besar. Bagaimana Anda sebaiknya mengidentifikasi video yang berkualitas baik berdasarkan tiga data numerik yang diberikan?**

Identifikasi video berkualitas baik dilakukan dengan melakukan filter video dengan views diatas rata-ratanya, likes diatas rata-ratanya, dan dislikes dibawah rata-ratanya karena video yang baik pastinya memiliki views dan likes sebanyak-banyaknya dan dislikes sesedikit-sedikitnya. Hasil akhirnya didapatkan bahwa jumlah video berkualitas baik ada sebanyak 1801.

```
In [69]: def good_quality_videos(df):
          views_threshold = df['views'].mean()
          likes_threshold = df['likes'].mean()
          dislikes_threshold = df['dislikes'].mean()
          passed_view_threshold = df[df['views'] >= views_threshold]
          passed_likes_threshold = passed_view_threshold[passed_view_threshold['likes'] >=
          passed_dislikes_threshold = passed_likes_threshold[passed_likes_threshold['disli
          return passed_dislikes_threshold
```

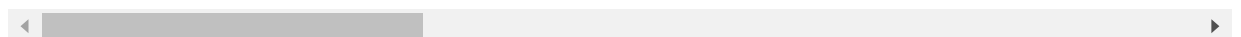
```
In [70]: good_videos = good_quality_videos(dataset)
          print("Banyak video dengan kualitas baik = %d" %(len(good_videos)))
          good_videos.head()
```

Banyak video dengan kualitas baik = 1801

```
Out[70]:
```

	video_id	trending_date	title	channel_title	category_id	publish_time	
63	ujyTQNNjjDU	17.14.11	G-Eazy - The Plan (Official Video)	GEazyMusicVEVO	10	2017-11- 10T05:00:01.000Z	BPG

	video_id	trending_date	title	channel_title	category_id	publish_time	
66	8mhTWqWlQzU	17.14.11	Wearing Online Dollar Store Makeup For A Week	Safiya Nygaard	22	2017-11-11T01:19:33.000Z	wea
98	jp9hK-jY6yY	17.14.11	When Someone Has A Crush on You   Lilly Singh	ISuperwomanII	23	2017-11-09T22:21:13.000Z	iisupe
105	g5c1bk8weaQ	17.14.11	FIRST TIME IM DOING THIS! TALKS WITH LIZA.	Liza Koshy	23	2017-11-10T03:43:43.000Z	
247	d380meD0W0M	17.15.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	r



Dari kumpulan video baik tersebut, kami memfilter lagi setiap video berdasarkan kategori ID yang ada. Mula-mula kami menghitung jumlah video di masing-masing kategori dan kemudian membuat histogram agar lebih mudah dianalisa. Kami juga mendapatkan nama-nama kategori dari website <https://techpostplus.com/youtube-video-categories-list-faqs-and-solutions/> [image.png](attachment:image.png). Setelah itu, kami dapat menyimpulkan bahwa video berkualitas baik didominasi oleh video dengan kategori musik.

```
In [71]: good_videos["category_id"].value_counts()
```

```
Out[71]: 10    563
         24    282
         22    234
         23    197
         26    151
          1    108
         28     73
         27     66
         20     61
         15     39
         17     24
          2      3
         Name: category_id, dtype: int64
```

```
In [72]: category = {
          "category_id": [1,2,10,15,17,19,20,22,23,24,25,26,27,28,29],
          "category_name": ["Film & Animation", "Autos & Vehicles","Music","Pets & Animals",
                           "Comedy","Entertainment","News & Politics","How to & Style","Ed
          }
```

```
category_df = pd.DataFrame.from_dict(category)
category_df
```

Out[72]:

	category_id	category_name
0	1	Film & Animation
1	2	Autos & Vehicles
2	10	Music
3	15	Pets & Animals
4	17	Sports
5	19	Travel & Events
6	20	Gaming
7	22	People & Blogs
8	23	Comedy
9	24	Entertainment
10	25	News & Politics
11	26	How to & Style
12	27	Education
13	28	Science & Technology
14	29	Nonprofits & Activism

In [73]:

```
good_videos_with_cat_name = pd.merge(good_videos, category_df, how = 'inner', on='category_id')
good_videos_with_cat_name.head()
```

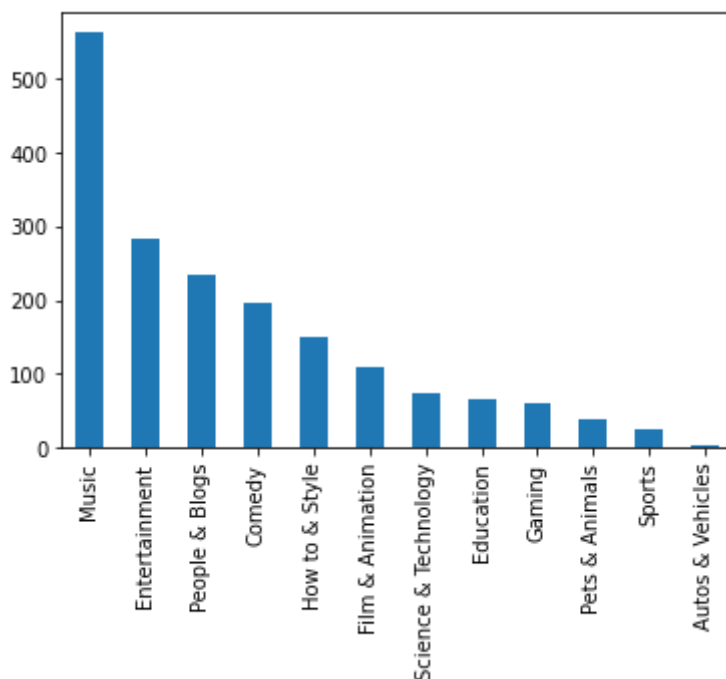
Out[73]:

	video_id	trending_date	title	channel_title	category_id	publish_time	
0	ujyTQNNjjDU	17.14.11	G-Eazy - The Plan (Official Video)	GEazyMusicVEVO	10	2017-11- 10T05:00:01.000Z	BPG/I
1	ujyTQNNjjDU	17.15.11	G-Eazy - The Plan (Official Video)	GEazyMusicVEVO	10	2017-11- 10T05:00:01.000Z	BPG/I
2	ixxR3ZoqnF0	17.17.11	BTS (방 탄소년 단) 'MIC Drop (Steve Aoki Remix)' Offi...	ibighit	10	2017-11- 16T15:00:05.000Z	

	video_id	trending_date	title	channel_title	category_id	publish_time	
3	BQ_0QLL2gqI	17.20.11	Hailee Steinfeld, Alesso - Let Me Go ft. Flori...	HaileeSteinfeldVEVO	10	2017-11-17T18:00:00.000Z	Hailee"S
4	BQ_0QLL2gqI	17.21.11	Hailee Steinfeld, Alesso - Let Me Go ft. Flori...	HaileeSteinfeldVEVO	10	2017-11-17T18:00:00.000Z	Hailee"S

```
In [74]: good_videos_with_cat_name['category_name'].value_counts().plot(kind='bar')
```

```
Out[74]: <AxesSubplot:>
```



## V. KESIMPULAN & SARAN

### Kesimpulan

- Video Youtube cenderung mendapat likes dibandingkan dislikes yang menunjukkan adanya algoritma Filter Bubble.
- Hanya ada 0.098% video viral dengan views > 100 juta.
- Untuk bisa dikategorikan terfavorit, video harus memiliki minimal 2.906.264 likes.
- Jika disaring dengan rata-rata views, likes, dan dislikes, terdapat 1801 video yang dapat dikatakan berkualitas baik yang didominasi dengan video musik.

### Saran

- Untuk memperjelas visualisasi data yang terlalu luas, dapat dilakukan zoom in.
- Salah satu cara untuk menentukan kualitas suatu video adalah dengan mengkategorikannya berdasarkan rata-rata views, likes, dan dislikes.

In [ ]: