

# **MATH1042 Laporan Proyek 1A - Kelompok 7**



## **Kelompok 7:**

- |           |                                     |                         |
|-----------|-------------------------------------|-------------------------|
| <b>1.</b> | <b>Christopher Jonathan</b>         | <b>212100170 / IBDA</b> |
| <b>2.</b> | <b>Dilivio Cullen Lemuel Tilaar</b> | <b>212100576 / IBDA</b> |
| <b>3.</b> | <b>Juan Christian Chandra</b>       | <b>212100108 / IBDA</b> |
| <b>4.</b> | <b>Petra William Leka</b>           | <b>212100331 / IEE</b>  |
| <b>5.</b> | <b>Wesley Hakim</b>                 | <b>212100211 / IEE</b>  |

**Calvin Institute of Technology**  
**Tahun ajaran 2022/2023**

## **Langkah-langkah Pengerjaan**

1. Mengeksplorasi isi dari dataset dan menentukan data yang akan digunakan.
2. Membuat rangkuman statistika kuantitatif dari tiga data parameter numerik, yaitu jumlah views, jumlah likes, dan jumlah dislikes. Rangkuman tersebut terdiri dari rata-rata, median, modus, standar deviasi, kuartil, jangkauan inter kuartil, dan pencilan dari masing-masing data parameter numerik yang dibuat menggunakan Python.
3. Membuat visualisasi data dari tiga data parameter numerik yang sebelumnya (dilakukan menggunakan Python). Pertama, membuat histogram untuk menunjukkan sebaran data secara komprehensif. Kemudian, membuat boxplot untuk melihat sebaran lokasi data dan posisi dari data-data pencilan. Histogram dibuat dengan log10 pada sumbu y frequency karena variasi nilai frequency terlalu besar.
4. Mendeskripsikan perilaku data jumlah views berdasarkan kelompok category\_id (dilakukan menggunakan Python). Dilakukan dengan menghitung rata-rata dari data jumlah views per kelompok category\_id. Kemudian, membuat histogram dan boxplot untuk perbandingan sebaran data jumlah views per kelompok category\_id.
5. Menjawab pertanyaan-pertanyaan diskusi dengan memakai prinsip distribusi standar normal.
6. Membuat PowerPoint untuk mempresentasikan hasil pekerjaan kami.
7. Membuat laporan.

## Intisari dan Visualisasi Data

1. Data yang diambil dan digunakan disini merupakan data dari kumpulan video youtube dari tahun ke tahun, dengan data yang disajikan sebanyak 40 ribuan video.
2. Informasi statistika umum yang mencakup Mean, Median, Modus, dan berbagai informasi lainnya pada data tersebut

- Mean

```
Rata-rata views = 2360784.6382573447  
Rata-rata likes = 74266.7024347359  
Rata-rata dislikes = 3711.400888910596
```

- Median

```
Median views = 681861.0  
Median likes = 18091.0  
Median dislikes = 631.0
```

- Modus

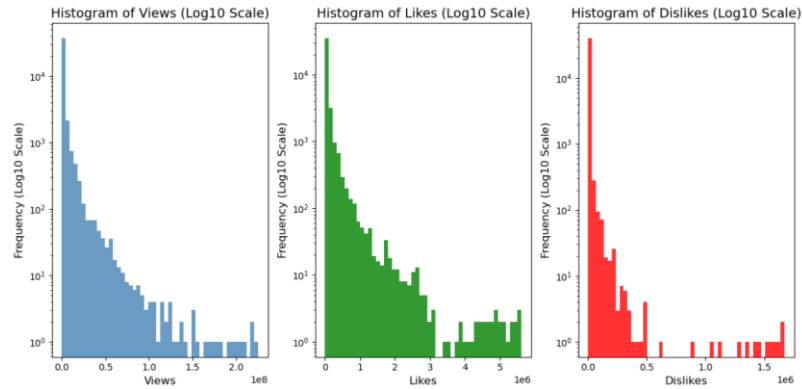
```
Modus views = 2078  
Modus likes = 0  
Modus dislikes = 0
```

- Standar Deviasi

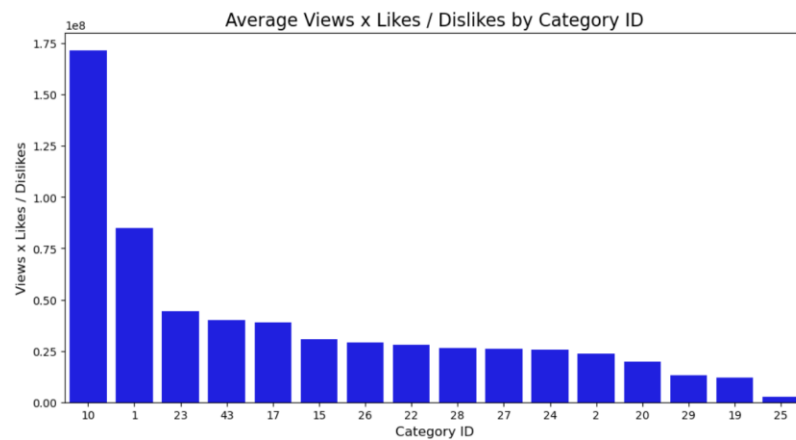
```
Standar deviasi views = 7394113.759703945  
Standar deviasi likes = 228885.33820949908  
Standar deviasi dislikes = 29029.70594500179
```

### 3. Grafik visualisasi distribusi data

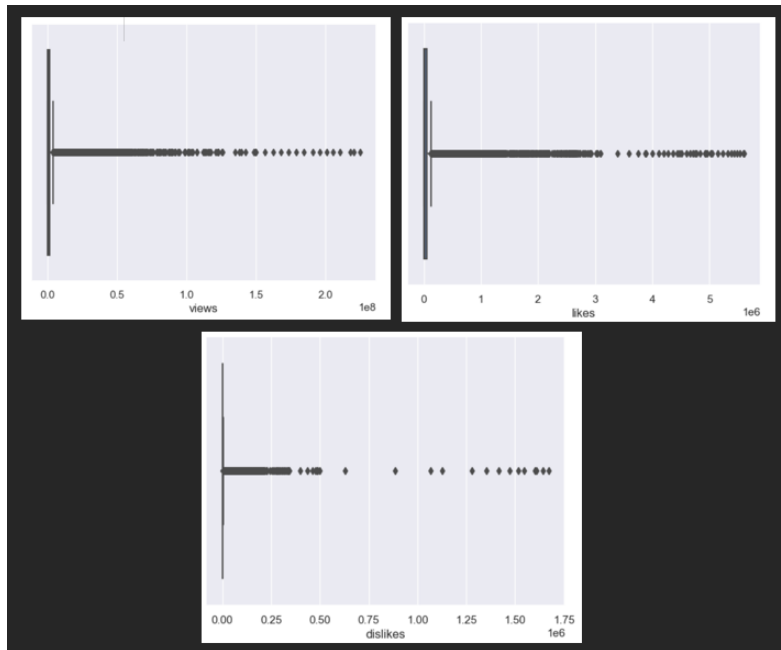
- Menurut views, likes, dan dislikes.



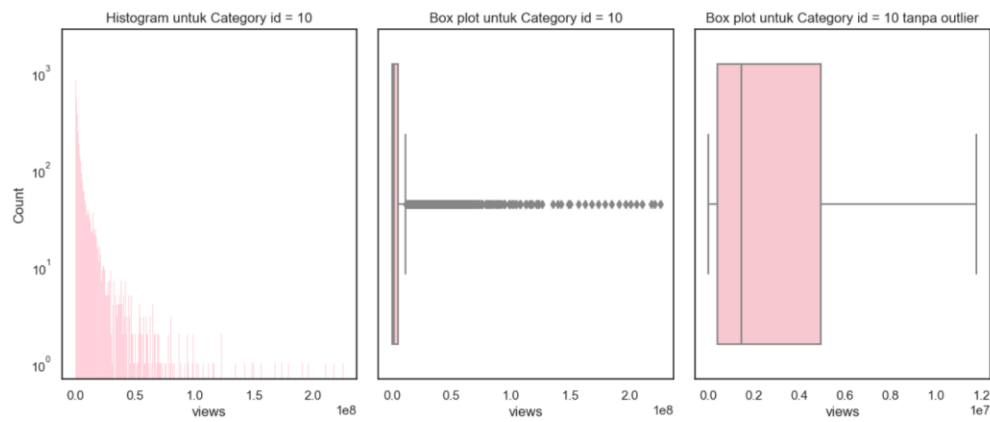
- Menurut rata-rata views x likes / dislikes pada setiap category ID



- Gambaran Pencilan menurut views, likes, dan dislikes



- Gambaran category dengan views terbanyak



## Deskripsi Pencilan Data

Untuk melihat pencilan data, data akan dibagi menjadi 4 ukuran kuantitatif *views*, *likes*, *dislikes*, dan juga *comment*.

- Views

○ Jumlah pencilan data	= 4499
○ Kuartil 1 (Q1)	= 242329
○ Kuartil 2 / median (Q2)	= 681861
○ Kuartil 3 (Q3)	= 1823157
○ IQR	= 1580828
○ Batas bawah pencilan ( $Q1 - 1.5 \cdot IQR$ )	= -2128913
○ Batas atas pencilan ( $Q3 + 1.5 \cdot IQR$ )	= 4194399

Dari data diatas terlihat bahwa terdapat 4499 video yang menjadi data pencilan dalam views. 4499 video ini semuanya berada di atas batas atas pencilan saja dikarenakan batas bawah pencilan bernilai negatif dan jumlah views tidak mungkin bernilai negatif.

- Likes

○ Jumlah pencilan data	= 5136
○ Kuartil 1 (Q1)	= 5424
○ Kuartil 2 / median (Q2)	= 18091
○ Kuartil 3 (Q3)	= 55417
○ IQR	= 49993
○ Batas bawah pencilan ( $Q1 - 1.5 \cdot IQR$ )	= -69565.5
○ Batas atas pencilan ( $Q3 + 1.5 \cdot IQR$ )	= 130406.5

Dari data diatas terlihat bahwa terdapat 5136 video yang menjadi data pencilan dalam likes. 5136 video ini semuanya berada di atas batas atas pencilan saja dikarenakan batas bawah pencilan bernilai negatif dan jumlah views tidak mungkin bernilai negatif.

- Dislikes

○ Jumlah pencilan data	= 5288
○ Kuartil 1 (Q1)	= 202
○ Kuartil 2 / median (Q2)	= 631
○ Kuartil 3 (Q3)	= 1938
○ IQR	= 1736
○ Batas bawah pencilan ( $Q1 - 1.5 \cdot IQR$ )	= -2402
○ Batas atas pencilan ( $Q3 + 1.5 \cdot IQR$ )	= 4542

Dari data diatas terlihat bahwa terdapat 5288 video yang menjadi data pencilan dalam dislikes. 5288 video ini semuanya berada di atas batas atas pencilan saja dikarenakan batas bawah pencilan bernilai negatif dan jumlah views tidak mungkin bernilai negatif.

- Comments

○ Jumlah pencilan data	= 5089
○ Kuartil 1 (Q1)	= 614
○ Kuartil 2 / median (Q2)	= 1856
○ Kuartil 3 (Q3)	= 5755
○ IQR	= 5141
○ Batas bawah pencilan ( $Q1 - 1.5 \cdot IQR$ )	= -7097.5
○ Batas atas pencilan ( $Q3 + 1.5 \cdot IQR$ )	= 13466.5

Dari data diatas terlihat bahwa terdapat 5089 video yang menjadi data pencilan dalam comments. 5089 video ini semuanya berada di atas batas atas pencilan saja dikarenakan batas bawah pencilan bernilai negatif dan jumlah views tidak mungkin bernilai negatif.

## Deskripsi Ekor Distribusi

- Deskripsi ekor distribusi pada views:

- Ekor distribusi kiri pada views adalah 549
- Ekor distribusi kanan pada views adalah 225211923

Views tersedikit dalam kumpulan video ini adalah 549 views dan views terbanyak dari kumpulan video ini adalah 225211923.

- Deskripsi ekor distribusi kiri pada likes:

- Ekor distribusi kiri pada likes adalah 0
- Ekor distribusi kanan pada like adalah 5613827

Likes tersedikit dalam kumpulan video ini adalah 0, dan menariknya modus dari likes dalam kumpulan video ini adalah 0 juga. Sedangkan likes terbanyak dari video ini adalah 5613827.

- Deskripsi ekor distribusi kiri pada dislike:

- Ekor distribusi kiri pada dislike adalah 0
- Ekor distribusi kanan pada dislike adalah 1674420

Dislike tersedikit dari kumpulan video ini adalah 0. Seperti likes, modus dari dislike juga merupakan 0. Sedangkan dislike terbanyak dari kumpulan video ini adalah 1674420.

- Deskripsi ekor distribusi kiri pada comment\_count:

- Ekor distribusi kiri pada comment\_count adalah 0
- Ekor distribusi kanan pada comment\_count adalah 1361580

Jumlah comment tersedikit dalam kumpulan video ini adalah 0, dan jumlah comment terbanyak dalam kumpulan video ini adalah 1361580.



## Ringkasan dan Interpretasi Data

Dari dataset di atas didapati bahwa persebaran views sangat menumpuk pada sekitar angka nol, yang berarti bahwa kemungkinan suatu video untuk dapat mencapai views yang banyak sangat sulit. Bisa dilihat juga bahwa dalam likes dan dislikes, data juga menumpuk pada angka nol sehingga bisa disimpulkan bahwa lebih banyak konsumen menonton saja tanpa berinteraksi dengan video tersebut seperti like dan dislike. Dapat dilihat juga bahwa konsumen cenderung untuk berinteraksi ketika ia menyukai video dibanding tidak menyukai video, berdasarkan data dimana jumlah like yang lebih banyak dibanding dislike pada histogram.

Untuk mencapai video yang termasuk berkualitas dan memiliki kemungkinan tinggi untuk ditonton, penting bagi pembuat video memperhatikan kategori video yang dibuat. Dari data yang didapatkan, kategori 10, 1, dan 23 terlihat memiliki nilai penonton yang lebih banyak sehingga bila pembuat video membuat video dalam kategori ini, video tersebut lebih mungkin untuk ditonton dan disukai dibandingkan kategori lain. Ini bisa disimpulkan dengan cara views rata-rata dikali likes rata-rata kemudian dibagi dislikes rata-rata untuk menunjukkan kemungkinan video tersebut ditonton dan disukai. Sehingga untuk membuat video yang berkualitas dan berkemungkinan tinggi untuk ditonton disarankan untuk membuat video dengan kategori 10.

### Highlights

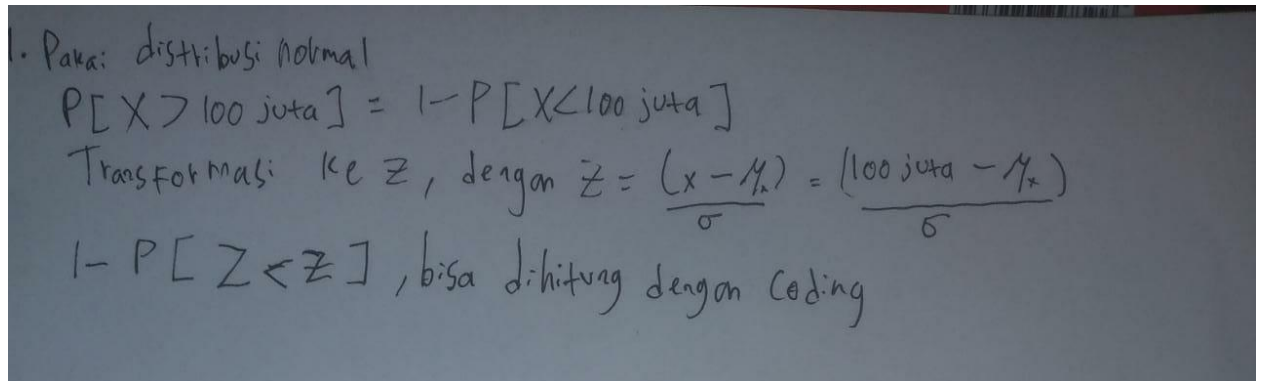
- Lebih banyak orang yang menonton tanpa berinteraksi dengan videonya.
- Orang-orang lebih cenderung berinteraksi saat dia menyukai video tersebut dibanding tidak menyukai.
- Persebaran views sangat menumpuk pada sekitar angka nol sehingga kemungkinan besar video kita tidak akan ditonton orang banyak.
- Pada kategori 10, 1, 23 terlihat memiliki nilai  $(\text{views rata-rata} * \text{likes rata-rata} / \text{dislikes rata-rata})$  penonton yang lebih banyak sehingga membuat video kategori ini lebih mungkin untuk ditonton dan disukai.
- Jika ingin membuat video termasuk kategori berkualitas dan memiliki kemungkinan ditonton tinggi, disarankan membuat video pada kategori 10.

## Pertanyaan Diskusi

1. Seberapa terpencil video youtube likes diatas 100 juta?
2. Berapa batas minimal likes agar hanya sekitar 0.1% video yang dapat tergolong video favorit?
3. Bagaimana cara mengidentifikasi video berkualitas berdasarkan data views, likes, dislikes?

Jawaban:

Untuk mendapat perilaku seluruh populasi dari sampel yang didapat, maka kami akan memakai Distribusi Normal Standar.



1. Pakai distribusi normal

$$P[X > 100 \text{ juta}] = 1 - P[X < 100 \text{ juta}]$$

Transformasi ke Z, dengan  $Z = \frac{(x - \mu)}{\sigma} = \frac{(100 \text{ juta} - \mu)}{\sigma}$

$1 - P[Z < Z]$ , bisa dihitung dengan coding

1.

Coding untuk mendapat hasil:

```
from scipy.stats import norm
import math
v = (10**8 - rata_rata[0])/std_deviasi[0]
print(rata_rata[0], std_deviasi[0], v)
print((1-norm.cdf(v, 0, 1)))
```

Hasil yang didapat adalah 0. Berarti, kemungkinan kita membuat video youtube dengan views diatas 100 juta adalah mendekati 0.

$$\begin{aligned}
 2. \quad & P[Z > z] = 0.001 \\
 & 1 - P[Z \leq z] = 0.001 \\
 & P[Z \leq z] = 0.999, \text{ pakai Coding} \\
 & z = 3.09, \text{ ubah } z \text{ ke } x \\
 & X = z \times \text{std} + \text{rata-rata} = 781576
 \end{aligned}$$

2.

Coding untuk mendapatkan hasil:

```

z = norm.ppf(0.999,0,1)
print(z)
print(f'Untuk bisa video like top 0.1% dibutuhkan like sebanyak
{z*std_deviasi[1]+rata_rata[1]}')

```

Outputnya adalah

“Untuk bisa video like top 0.1% dibutuhkan like sebanyak 781575.5689778763”.

Maka diperkirakan dibutuhkan 781576 likes untuk menjadi video favorit.

3. Untuk menemukan video berkualitas, kami mengambil video yang memiliki likes top 0.1% dibandingkan video lain sekaligus perbandingan likes dan dislikes adalah 75:1.

Code yang dipakai adalah:

```

#syarat kualitas baik video favorit sekaligus perbandingan antara
likes/dislikes = 75
syt = ( df['likes'] >= 781575.5689778763 ).sum()
print(syt)

df['like_dislike_ratio'] = df.apply(lambda row: row['likes'] /
row['dislikes'] if row['likes'] >= 781576 else np.nan, axis=1)
# jumlah data yang perbandingan likes dan dislikes-nya kurang dari
0.75
jumlah_kurang_dari = len(df.loc[df['like_dislike_ratio'] < 75])
# jumlah data yang perbandingan likes dan dislikes-nya lebih dari
0.75
jumlah_lebih_dari = len(df.loc[df['like_dislike_ratio'] > 75])
print(f'Dari {syt} video favorit, hanya {jumlah_lebih_dari} video
yang memenuhi syarat video berkualitas')

```

Outputnya adalah “Dari 540 video favorit, hanya 127 video yang memenuhi syarat video berkualitas”.