Implementasi Teknik *Web Scraping* dan Analisis *Clustering* pada Data Top 500 IMDb US *Box Office Movies*

Azka Muhammad Radinka Purba[1*], Aubert Oktaviantono[2], Adatul Mukarommah[3]
[1, 2, 3]Statistika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*Corresponding author e-mail: azkaradinka@gmail.com

***Abstrak*—Perkembangan teknologi yang semakin pesat membuat proses pencarian dan pengambilan data dalam jumlah besar dengan lebih mudah didapat dan cepat. Data bisa didapatkan dari halaman website dengan metode *web scraping. Web scraping* adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman website. Penulis ingin mengkaji data 500 film teratas versi IMDb menggunakan *web scraping,* eksplorasi data, dan *K-Means Clustering* untuk mengetahui informasi dan karakteristik dari film-film yang masuk 500 film teratas versi IMDb. Dari analisis diketahui film dengan genre Drama menjadi yang paling banyak masuk dalam 500 film teratas IMDb. Film dengan kategori capaian rating yang Sangat Baik cenderung memiliki rata-rata Runtime dan Gross yang tinggi dibandingkan dengan capaian rating lain. Dengan analisis *cluster* menggunakan *K-Means Clustering* didapatkan bahwa terdapat 401 film yang terkelompok di *cluster* 1, dan 99 film terkelompok menjadi *cluster* 2. *Cluster* 1 memiliki nilai median dan mean yang lebih rendah untuk semua variabel dibandingkan dengan *cluster* 2. Penulis menyarankan agar penelitian selanjutnya bisa melakukan eksplorasi data variabel lain dan eksplorasi metode analisis *cluster* lain yang melibatkan variabel-variabel kategorik sebagai variabel penentuan *cluster*. Penulis merekomendasikan IMDb agar mengumpulkan kritik dan review film dari kritikus-kritikus terpercaya agar dapat melengkapi data Metascore pada film-film yang belum memiliki Metascore.**

***Kata Kunci*— *Average Silhouette Width*, Eksplorasi Data, Elbow Plot, IMDb, *K-Means Clustering*, *Web Scraping*.**

I. PENDAHULUAN

Seiring dengan perkembangannya teknologi, proses pencarian data yang diperlukan untuk keperluan riset dan penelitian semakin mudah didapat. Data menjadi bisa didapatkan dimana saja. Data tidak selalu harus didapat dengan proses sensus ataupun survei. Saat ini data bisa kita dapatkan dari website yang sekarang tersebar di internet. Salah satu metode yang bisa digunakan untuk mengambil data dari website adalah *web scraping*.

Web scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman website dalam bahasa markup seperti HTML (*HyperText Markup Language*) atau XHTML (*Extensible HyperText Markup Language*), dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain [1].

Data yang sudah didapatkan menggunakan metode *web scraping* bisa di manipulasi menjadi data yang sesuai dengan kebutuhan kita dalam analisis dan visualisasi data. Data yang sebelumnya hanya menampilkan sedikit informasi, selanjutnya dapat di eksplorasi untuk memberikan informasi yang lebih banyak dan lebih bermanfaat untuk kebutuhan penelitian dan pembuatan keputusan.

Dalam analisis ini, penulis ingin melakukan *web scraping* dan analisis *cluster* menggunakan metode *k-means clustering* untuk data dari 500 film teratas di website IMDb. Analisis *cluster* merupakan seperangkat metode yang digunakan untuk mengelompokkan objek ke dalam sebuah *cluster* berdasarkan informasi yang ditemukan pada data. Hasil *cluster* dikatakan baik ketika mempunyai homogenitas yang besar antar objek dalam satu *cluster* dan heterogenitas yang besar pula antar *cluster* yang satu dengan *cluster* lainnya [2]. IMDb adalah situs resmi populer yang menyediakan informasi mengenai film, acara televisi, selebriti. IMDb adalah produk dan jasa untuk membantu penggemar dalam menentukan film atau acara apa yang ingin ditonton [3].

Penulis melakukan *web scraping* untuk pengambilan data 500 film teratas di website IMDb dan melakukan analisis *K-means clustering* menggunakan bahasa pemrograman R dalam aplikasi RStudio. Metode *K-Means Clustering* dipilih dalam analisis ini karena penulis ingin mengelompokkan data *Top 500 IMDb US Box Office Movies* menjadi beberapa label pengelompokkan yang jelas dari jumlah pendapatan kotor, nilai Rating, nilai Metascore, durasi film, dan jumlah vote yang didapatkan. Selain itu metode *K-means clustering* merupakan metode yang umum untuk analisis *clustering*.

Analisis menggunakan metode *web scraping* dan *K-means clustering* bertujuan sebagai media belajar penulis dalam menggunakan metode *web scraping* dan *K-means clustering* untuk mengambil data dari internet dan mengelompokkan data 500 film teratas di website IMDb menjadi beberapa *cluster* dengan karakteristik tertentu berdasarkan variabel jumlah pendapatan kotor, nilai Rating, nilai Metascore, durasi film, dan jumlah vote yang didapatkan. Selain itu analisis ini juga menjadi media belajar penulis dalam membuat *dashboard* dan *markdown* sebagai media untuk menampilkan hasil dari eksplorasi data dan analisis *cluster* dalam analisis ini.

II. TINJAUAN PUSTAKA

A. Statistika Deskriptif

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus dat sehingga memberikan informasi yang berguna. Statistika deskriptif merupakan bagian dari statistika berkaitan dengan cara meringkas data dalam ukuran-ukuran tertentu yang berbentuk tabel, diagram, grafik, dan besaran-besaran lain [4].

Hasil ukuran pemusatan data dapat dijadikan pedoman untuk mengamati karakter dari sebuah data. Ukuran pemusatan data dapat berupa rata-rata, median, modus, kuartil bawah, dan kuartil atas. Ukuran penyebaran data digunakan untuk menentukan seberapa besar nilai-nilai data berbeda atau bervariasi dengan nilai pusatnya, atau seberapa besar data tersebut menyimpang dari nilai pusatnya. Ukuran penyebaran data terdiri dari: jangkauan, variasi, dan standar deviasi [4].

B. Eksplorasi Data

Analisis Eksplorasi data adalah proses menggali intuisi (memahami) terhadap data yang digunakan, apakah data ini dan untuk apa data ini [5]. Beberapa plot yang bisa digunakan sebagai berikut.

1) Barchart

Bar chart adalah grafik dengan batangan persegi panjang, yang biasanya digunakan untuk membandingkan kategori-kategori yang berbeda. Sumbu horizontal menampilkan kategori, dan sumbu vertikal menampilkan nilai dari kategori tersebut, biasanya nilai jumlah data, atau persentase [7].

2) Histogram

Histogram adalah tumpukan kolom persegi panjang yang panjangnya proporsional terhadap jumlah nilai (dibagi dengan lebar dasar, jika kolom tidak memiliki lebar yang sama) [8]. Histogram dibentuk dengan menempatkan variabel yang diminati pada sumbu horizontal dan frekuensi, frekuensi relatif atau persen distribusi frekuensi pada sumbu vertikal.

3) Radar Chart

Radar chart adalah metode grafis untuk menampilkan data multivariat dalam bentuk grafik dua dimensi dari tiga atau lebih variabel kuantitatif. Variabel direpresentasikan dengan sumbu dari titik yang sama. Posisi relatif dan sudut sumbu tidak informatif [9].

4) Scatterplot

Scatterplot adalah grafik yang berisi variabel independen di sepanjang sumbu x horizontal dan variabel dependen di sepanjang sumbu y vertikal. Setiap titik di sebar mewakili satu kasus [10]. Scatterplot yang biasa digunakan untuk melihat suatu pola hubungan antara 2 variabel. Untuk menggunakan scatterplot skala data yang digunakan haruslah skala interval dan rasio.

C. Web Scraping

Web scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman website dalam bahasa markup seperti HTML (*HyperText Markup Language*) atau XHTML (*Extensible HyperText Markup Language*), dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain [1].

D. Metode Elbow

Dalam metode elbow, varians (*within-cluster sum of square errors*) diplotkan terhadap jumlah *cluster*. Beberapa *cluster* pertama akan menunjukkan banyak variasi dan informasi, tetapi pada titik tertentu, perolehan informasi akan menjadi rendah, sehingga memberikan struktur sudut ke grafik. Jumlah *cluster* optimal ditemukan dari titik ini; oleh karena itu, ini dikenal sebagai "*elbow criterion*". Tetapi poin ini tidak selalu dapat ditentukan tanpa rasa ambiguitas [11].

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \left| |X_i - C_k| \right|_2^2 \quad (1)$$

E. Metode Average Silhouette

Dalam metode *average silhouette width*, nilai siluet untuk setiap titik data dihitung, yang meannya digunakan untuk menemukan jumlah *cluster* yang optimal. Nilai *silhouette* menunjukkan kemiripan titik data dengan *cluster*nya sendiri jika dibandingkan dengan semua *cluster* atau pusat *cluster* lainnya. Nilainya berkisar dari -1 hingga +1. Nilai siluet yang lebih tinggi menyiratkan bahwa titik data cocok dengan pusat / *cluster*nya sendiri dan tidak begitu cocok dengan *cluster* lain. Jika mean dari nilai silhouette yang diukur untuk semua titik data cukup tinggi, maka dapat dikatakan jumlah *cluster* berada pada nilai optimalnya, atau dengan kata lain struktur *cluster*ing sudah sesuai. Di sisi lain, jika nilai siluet rata-rata ternyata sangat kurang atau negatif, maka itu berarti struktur *cluster* tidak tepat, dan mungkin memiliki jumlah *cluster* lebih banyak atau lebih sedikit dari nilai optimal. Untuk menemukan nilai siluet, metrik jarak apa pun, seperti jarak Minkowski atau jarak Euclidean, dapat digunakan [11].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

F. Algoritma K-Means Clustering

K-means merupakan salah satu metode data *cluster*ing non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* / kelompok. Metode ini mempartisi ke dalam *cluster* / kelompok sehingga data yang memiliki karakteristik yang sama (*High intra class similarity*) dikelompokkan ke dalam satu *cluster* yang sama dan yang memiliki karakteristik yang berbeda (*Low inter class similarity*) dikelompokkan pada kelompok yang lain [13]. Proses *cluster*ing dimulai dengan mengidentifikasi data yang akan di*cluster*, X_{ij} $(i = 1, \dots, \mathrm{n};\ \mathrm{j} = 1, \dots, \mathrm{m})$ dengan n adalah jumlah data yang akan di*cluster* dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap *cluster* ditetapkan secara bebas (sembarang), $C_{kj}(\mathrm{k} = 1, \dots, \mathrm{k};\ \mathrm{j} = 1, \dots, \mathrm{m})$. Kemudian dihitung jarak antara setiap data dengan setiap pusat *cluster*. Untuk melakukan penghitungan jarak data ke-i (x_i) pada pusat *cluster* ke-k (c_k), diberi nama (d_{ik}), dapat digunakan formula *Euclidean* [14] seperti pada persamaan berikut.

$$d_{ik} = \sqrt{\left(\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2\right)} \quad (3)$$

Suatu data akan menjadi anggota dari *cluster* ke-k apabila jarak data tersebut ke pusat *cluster* ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat *cluster* lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (4). Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap *cluster*.

$$Min \sum_{k=1}^{k} d_{ik} = \sqrt{\left(\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2\right)} \quad (4)$$

Nilai pusat *cluster* yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada *cluster* tersebut, dengan menggunakan rumus pada persamaan berikut.

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (5)$$

Dimana $x_{ij} \in$ *cluster* ke – k dan p = banyaknya anggota *cluster* ke k.

Algoritma dasar dalam *k-means clustering* adalah sebagai berikut.

1. Tentukan jumlah *cluster* (k), tetapkan pusat *cluster* sembarang.
2. Hitung jarak setiap data ke pusat *cluster* menggunakan jarak *euclidean*.
3. Kelompokkan data ke dalam *cluster* berdasarkan jarak terpendek.
4. Hitung pusat *cluster* menggunakan nilai rata-rata dari data yang sudah ter*cluster*.
5. Ulangi langkah 2 sampai dengan 4 hingga sudah tidak ada lagi data yang berpindah ke *cluster* yang lain.

III. METODE ANALISIS

A. *Sumber Data*

Dalam analisis ini, data yang digunakan adalah data IMDB Top 500 Film yang diurutkan berdasarkan US Box Office. Data diambil pada tanggal 25 Desember 2020 pukul 20.00 WIB.

B. *Variabel Analisis*

Variabel yang digunakan dalam analisis terdiri dari lima variabel yakni sebagai berikut:

Tabel. 1.
Tabel Variabel Analisis

No	Variabel	Deskripsi
X1	Title	Judul film
X2	Rating	Penilaian penonton
X3	Gross	Pendapatan kotor film
X4	Metascore	Penilaian kritikus film
X5	Runtime	Durasi film
X6	Director	Sutradara film
X7	Year	Tahun rilis film

C. *Library Bahasa Pemrograman*

Library Pemrograman R yang digunakan dalam analisis ini yakni sebagai berikut:

Tabel. 2.
Tabel *Library* Pemrograman

Library	Kegunaan
readxl	Membaca data dengan format .xlsx
writexl	Membuat file .xlsx
xml2	*Scrapping* data
rvest	*Scrapping* data
stringr	*String manipulation*
rebus	*String manipulation*
BBmisc	*Data Wrangling*
purrr	*Data Wrangling*
data.table	*Data Wrangling*
dplyr	Pengolahan data
ggplot2	Visualisasi data
plotly	Visualisasi data
cowplot	Visualisasi data
grDevices	Visualisasi data
ggpubr	Visualisasi data
fmsb	Visualisasi data
formattable	Visualisasi data
radarchart	Visualisasi data
DT	Visualisasi data
dygraph	Visualisasi data
knitr	Membuat dashboard
shiny	Membuat dashboard
shinydashboard	Membuat dashboard
flexdashboard	Membuat dashboard
prettydoc	Membuat Rmarkdown
factoextra	*Clustering*

D. *Langkah Analisis*

Langkah analisis yang digunakan dalam analisis ini yakni sebagai berikut:

1) Scrapping data
   a. Mencari halaman web yang akan di *scrapping*.
   b. Mendapatkan bahasa HTML data yang ingin kita *scrapping* dari halaman website.
   c. Melakukan *scrapping* data
   d. Menggabungkan hasil *scrapping* menjadi satu *dataframe*.
2) Pra-pemrosesan data
3) Eksplorasi Data
4) Penentuan jumlah *cluster* untuk analisis *clustering* menggunakan :
   a. Metode *Elbow*
   b. Metode *Average Silhouette*
5) Analisis *cluster*ing
6) Menarik kesimpulan dan saran

IV. HASIL DAN PEMBAHASAN

A. *Scrapping data*

Scrapping dilakukan pada website IMDB pada link https://www.imdb.com/search/title/?groups=top_1000&sort=boxoffice_gross_us,desc. Tahap pertama adalah mengambil nodes variabel-variabel target pada halaman html website menggunakan CSS *selector extension*

google chrome yaitu selectorgadet. Setelah nodes untuk setiap variabel target diperoleh, selanjutnya dilakukan pengekstrakkan data dari nodes html menggunakan library rvest pada *software* R. Penulis hanya mengambil 500 data karena untuk data 500 sampai 1000 didapatkan banyak film yang belum memiliki metascore sehingga penulis hany membatasi sampai 500 data. Pada link tersebut data 500 film terpisah menjadi 10 halaman sehingga dilakukan *looping* dari halaman 1 hingga 10. Setelah seluruh variabel target sudah diekstrak, kemudian data-data variabel tersebut disatukan dan dijadikan dataframe mentah yang siap untuk dibersihkan ditahap selanjutnya.

B. *Pra-Pemrosesan Data*

Pada tahap ini terdapat beberapa hal yang akan dilakukan diantaranya adalah :

1. Membersihkan *space* dan tab yang terdapat pada kolom title.
2. Membuat variabel tahun dengan mengekstrak dari kolom title menggunakan *regular expression*.
3. Membersihkan *space* dan tab yang terdapat pada kolom synopsis.
4. Mengubah tipe data votes menjadi tipe data numerik dengan pemisah desimal menggunakan titik.
5. Membersihkan kolom gross dari lambang dollar dan huruf M, dan mengubahnya ke tipe data numerik
6. Membersihkan kolom runtime dengan menghapus *space* dan kata “min”, lalu mengubahnya ke tipe data numerik.
7. Memisahkan kolom genre film menjadi 3 kolom dikarenakan data genre film yang didapat bervariasi dari 1 genre hingga 3 genre untuk 1 film.

```
$ Title             <chr> "\n        1.\
$ Rating            <chr> "7.9", "8.4", "
$ Synopsis          <chr> "\n    As a ne
$ Votes             <chr> "857,241", "79
$ Gross             <chr> "$936.66M", "$
$ Runtime           <chr> "138 min", "18
$ Genre             <chr> "\nAction, Adv
$ Director          <chr> "J.J. Abrams",
$ Director_and_stars <chr> "\n    Directo
```

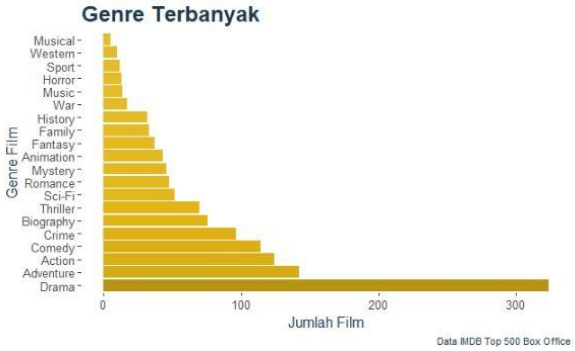
Gambar. 1. Raw dataframe

Setelah dilakukan pembersihan data, didapatkan dataframe baru sebagai berikut.

```
Rows: 500
Columns: 16
$ Title      <chr> "Star Wars: Episode VII - The
$ Rating     <dbl> 7.9, 8.4, 7.8, 8.4, 7.8, 8.0,
$ Synopsis   <chr> "\r\n    As a new threat to t
$ Votes      <dbl> 856017, 792722, 1113972, 822
$ Gross      <dbl> 936.6, 858.3, 760.5, 678.8, 6
$ Runtime    <dbl> 138, 181, 162, 149, 194, 143,
$ Director   <chr> "J.J. Abrams", "Anthony Russo
$ Year       <dbl> 2015, 2019, 2009, 2018, 1997,
$ Genre_1    <chr> "Action", "Action", "Action",
$ Genre_2    <chr> "Adventure", "Adventure", "Ac
$ Genre_3    <chr> "Sci-Fi", "Drama", "Fantasy",
$ Stars_1    <chr> "Daisy Ridley", "Robert Downe
$ Stars_2    <chr> "John Boyega", "Chris Evans",
$ Stars_3    <chr> "Oscar Isaac", "Mark Ruffalo"
$ Stars_4    <chr> "Domhnall Gleeson", "Chris He
$ Metascore  <dbl> 80, 78, 83, 68, 75, 69, 80, 8
```

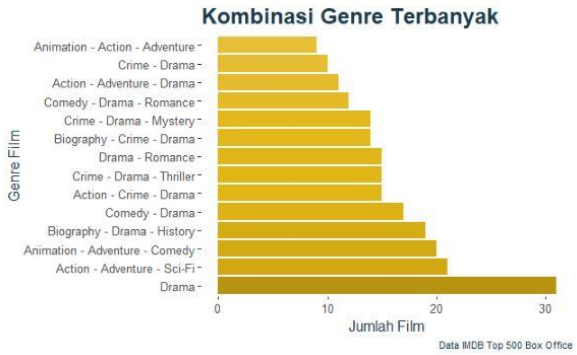
Gambar. 2. Clean dataframe

C. *Eksplorasi Data*



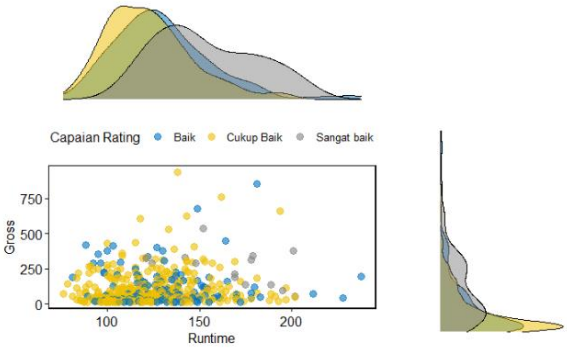
Gambar. 3. Genre terbanyak

Gambar 3 menunjukkan bar chart untuk jumlah film pada setiap genre film yang masuk kedalam top 500 IMDB *Box Office*. Dapat diketahui bahwa genre drama merupakan genre film yang paling banyak masuk kedalam top 500 IMDB *Box Office*. Selisih genre drama sebagai genre terbanyak dengan genre *adventure* yang berada di posisi kedua berbeda cukup jauh yakni 2 kali lipatnya.



Gambar. 4. Kombinasi genre terbanyak

Gambar 4. menunjukkan bar chart untuk jumlah film pada setiap kombinasi genre film yang masuk kedalam top 500 IMDB *Box Office*. Dapat diketahui bahwa film yang bergenre drama tanpa kombinasi dengan genre lain merupakan film yang paling banyak masuk kedalam top 500 IMDB *Box Office*.



Gambar. 5. Scatterplot dan histogram runtime vs gross

Gambar 5 menunjukkan sebaran data variabel runtime dan gross untuk setiap capaian rating yang terbagi atas capaian cukup baik (rating < 8), capaian baik (rating 8-8.5), dan capaian rating sangat baik (8.5-10). Dapat dilihat dari histogram yang berada di atas dan di sebelah kanan scatterplot bahwa film dengan rating yang sangat baik

memiliki sebaran data dengan gross dan runtime yang lebih tinggi jika dibandingkan dengan sebaran data film dengan capaian cukup baik dan baik. Dapat dikatakan film-film dengan capaian rating yang sangat baik memiliki rata-rata gross yang tinggi.

Tabel. 3.
Tabel Top Director

Director	Jumlah Film
Steven Spielberg	14
Christopher Nolan	8
Clint Eastwood	8
David Fincher	8
Martin Scorsese	7
Quentin Tarantino	7

Tabel 3 menunjukkan 6 director paling sukses dengan melihat jumlah film yang mereka sutradarai yang berhasil masuk di Top 500 IMDB *Box Office.* Selanjutnya akan dibentuk visualisasi radar chart pada 3 director dengan jumlah film paling banyak, untuk melihat dan memetakan rating, metascore, gross dan votes hasil dari film yang mereka sutradarai.

Title	Rating	Metascore	Gross	Votes
E.T. the Extra-Terrestrial	7.8	91	435.10	369985
Jurassic Park	8.1	68	402.40	861911
Jaws	8.0	87	260.00	541467
Raiders of the Lost Ark	8.4	85	248.10	879084
Saving Private Ryan	8.6	91	216.50	1226702
Indiana Jones and the Last Crusade	8.2	65	197.10	688410
Indiana Jones and the Temple of Doom	7.6	57	179.80	446445
Catch Me If You Can	8.1	75	164.60	821657
Close Encounters of the Third Kind	7.6	90	132.00	183822
Minority Report	7.6	80	132.00	505856
The Color Purple	7.8	78	98.47	77848
Schindler's List	8.9	94	96.90	1205371
Bridge of Spies	7.6	81	72.31	285737
Empire of the Sun	7.7	62	22.24	115014

Gambar. 6. Htmlwidget formattable film Steven Spielberg

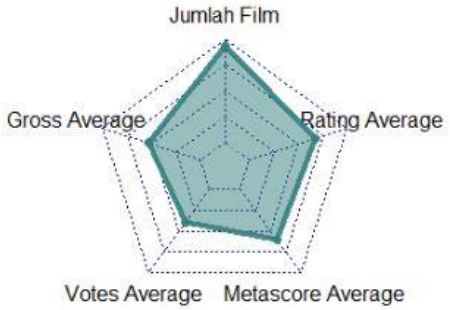
Title	Rating	Metascore	Gross	Votes
The Dark Knight	9.0	84	534.80	2286617
The Dark Knight Rises	8.4	78	448.10	1506521
Inception	8.8	74	292.50	2050183
Batman Begins	8.2	70	206.80	1300110
Dunkirk	7.9	94	188.30	548736
Interstellar	8.6	74	188.00	1495086
The Prestige	8.5	66	53.09	1180778
Memento	8.4	80	25.54	1118631

Gambar. 7. Htmlwidget formattable film Christopher Nolan

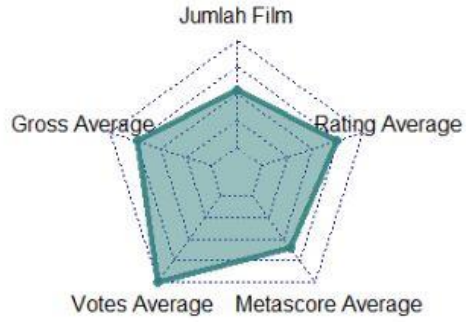
Title	Rating	Metascore	Gross	Votes
Gran Torino	8.1	72	148.10	716934
Unforgiven	8.2	85	101.10	373646
Million Dollar Baby	8.1	86	100.40	632469
Mystic River	7.9	84	90.14	417084
The Bridges of Madison County	7.6	69	71.52	72670
Changeling	7.7	63	35.74	238058
The Outlaw Josey Wales	7.8	69	31.80	65280
Letters from Iwo Jima	7.9	89	13.76	153535

Gambar. 8. Htmlwidget formattable film Christopher Nolan

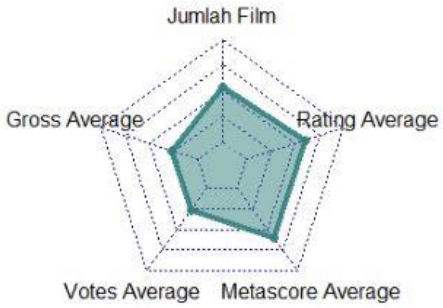
Gambar 6 sampai 8 menunjukkan data film yang disutradai oleh 3 top director. Digunakan formattable untuk melihat secara kasar perbandingan masing-masing nilai dalam variabel antar film.



Gambar. 9. Radar chart Steven Spielberg



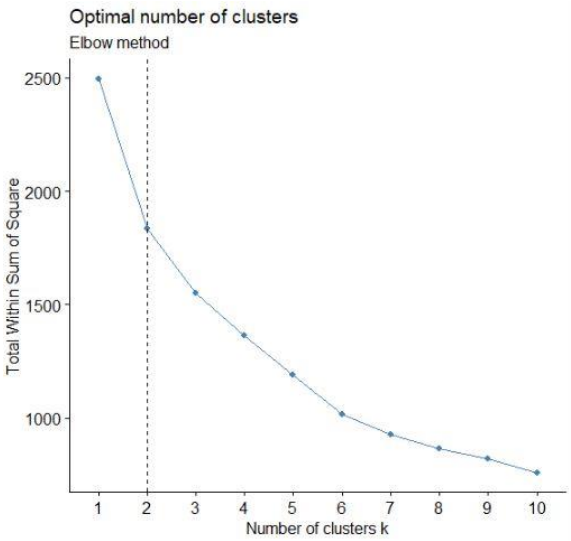
Gambar. 10. Radar chart Christopher Nolan



Gambar. 11. Radar chart Clint Eastwood

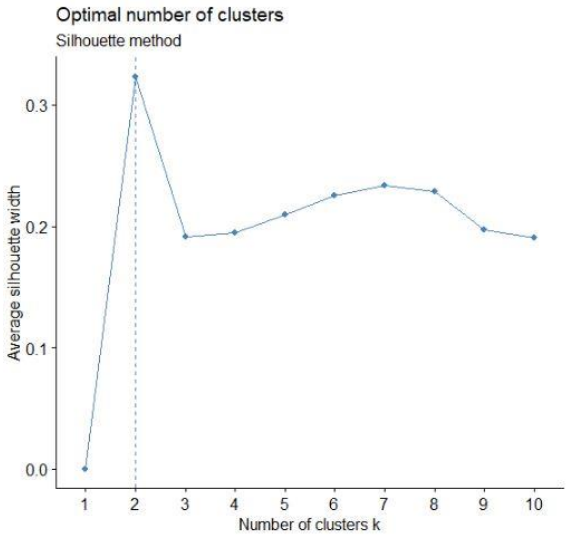
Gambar 9 sampai 11 merupakan radar chart 2 dimensi yang berisi 5 variabel kuantitatif yang direpresentasikan pada sumbu yang dimulai dari titik yang sama. Kedua director diatas merupakan director terbaik di dunia, namun kedua director memiliki capaian yang berbeda dalam hal jumlah film dan votes average. Steven Spielberg memiliki jumlah film yang lebih banyak yang masuk ke Top 500 IMDB Box Office, sementara Christopher Nolan walaupun jumlah filmnya tidak sebanyak Steven Spielberg, Christopher Nolan memiliki angka votes average yang lebih tinggi.

D. Penentuan jumlah cluster



Gambar. 12. Elbow Plot

Gambar 12 menunjukkan hasil penentuan jumlah *cluster* menggunakan metode Elbow. Dari gambar diatas dapat diketahui bahwa nilai yang optimal untuk jumlah *cluster* adalah 2, terlihat dari *elbow* atau patahan garis yang berada di $k = 2$.



Gambar. 13. Plot Silhouette Method

Gambar 13 menunjukkan hasil penentuan jumlah *cluster* menggunakan metode Average Silhouette Width. Dari gambar diatas dapat diketahui nilai optimal untuk jumlah *cluster* adalah 2, terlihat dari nilai *average silhouette width* tertinggi di $k = 2$. Sehingga dalam analisis ini, digunakan 2 *cluster* untuk analisis *clustering*.

E. K-Means Clustering

Proses *K-Means Clustering* dilakukan dengan *scaling* data. Selanjutnya data yang sudah di *scaling* dimasukkan ke dalam fungsi kmeans() untuk menentukan *cluster* setiap data.

```
12
13  df=read_xlsx('Data.xlsx')
14
15  data=df %>%
16    select(Rating, Votes, Gross, Runtime, Metascore)
17
18  # Scaling
19  data_scaled=scale(data)
20
21  model=kmeans(data_scaled, centers=2)
22  data_clustered = mutate(data, cluster=model$cluster)
23
24  head(data_clustered)
25
```

Gambar. 14. Syntax K-Means *Cluster*ing

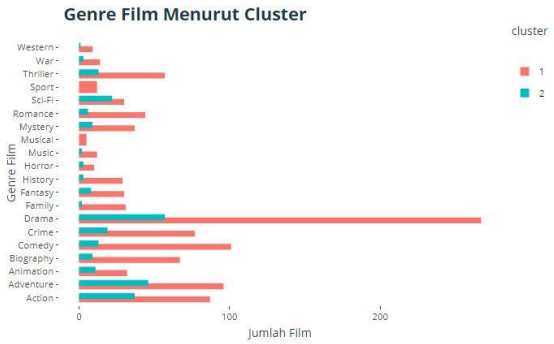
```
     Rating           Votes            Gross            Runtime         Metascore        cluster1
 Min.   :7.600   Min.   :  27548   Min.   :  10.63   Min.   : 76.0   Min.   : 28.00   Length:401
 1st Qu.:7.700   1st Qu.: 139914   1st Qu.:  27.30   1st Qu.:105.0   1st Qu.: 69.00   Class :character
 Median :7.800   Median : 262617   Median :  52.93   Median :120.0   Median : 77.00   Mode  :character
 Mean   :7.827   Mean   : 303073   Mean   :  84.57   Mean   :122.7   Mean   : 75.98
 3rd Qu.:8.000   3rd Qu.: 442601   3rd Qu.: 111.50   3rd Qu.:135.0   3rd Qu.: 84.00
 Max.   :8.500   Max.   :798435    Max.   :608.50    Max.   :212.0   Max.   :100.00
```

Gambar. 15. Summary statistik untuk setiap *cluster*

```
     Rating           Votes            Gross            Runtime         Metascore        cluster2
 Min.   :7.800   Min.   : 267189   Min.   :  13.09   Min.   :  81.0   Min.   : 57.00   Length:99
 1st Qu.:8.100   1st Qu.: 756660   1st Qu.:  77.25   1st Qu.:116.5   1st Qu.: 69.00   Class :character
 Median :8.300   Median : 935351   Median : 183.60   Median :136.0   Median : 80.00   Mode  :character
 Mean   :8.363   Mean   :1011133   Mean   : 221.46   Mean   :138.6   Mean   : 79.19
 3rd Qu.:8.500   3rd Qu.:1181951   3rd Qu.: 312.30   3rd Qu.:152.5   3rd Qu.: 89.00
 Max.   :9.300   Max.   :2333484   Max.   : 936.60   Max.   :238.0   Max.   :100.00
```

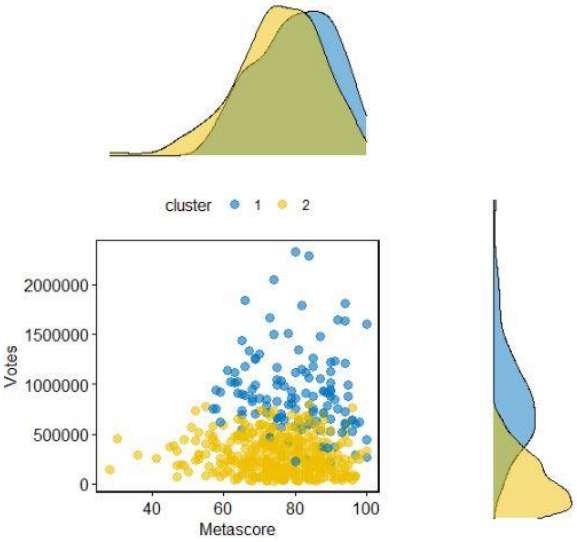
Gambar. 16. Summary statistik untuk setiap *cluster*

Gambar 15 dan 16 menunjukkan summary statistik setiap variabel untuk setiap *cluster* yang terbentuk. Dapat diketahui bahwa terdapat 401 film yang terkelompok di *cluster* 1, dan 99 film terkelompok menjadi *cluster* 2. *Cluster* 1 memiliki nilai median dan mean yang lebih rendah untuk semua variabel dibandingkan dengan *cluster* 2. Untuk mengetahui lebih dalam mengenai perbedaan karakteristik pada kedua *cluster*, digunakan visualisasi bar chart dan scatterplot untuk melihat karakteristik variabel-variabel yang digunakan.



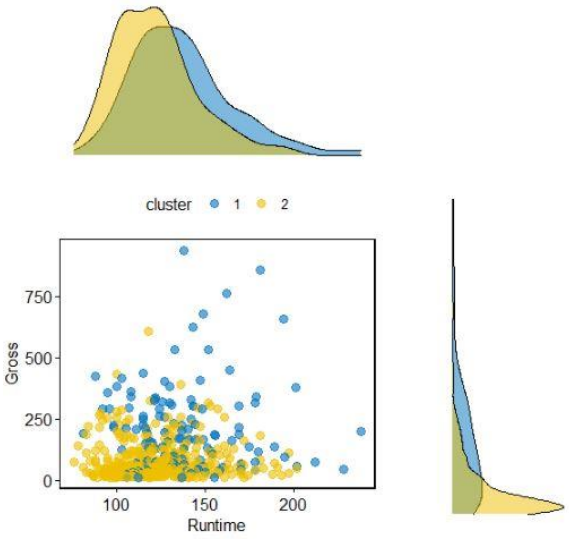
Gambar. 17. Bar chart Genre film untuk setiap *cluster*

Gambar 17 menunjukkan visualisasi bar chart jumlah genre film berdasarkan *cluster*. Dapat diketahui bahwa film dengan genre sport dan musical masuk di *cluster* 1, sementara film dengan genre lainnya tersebar ke kedua *cluster*. Diketahui perbedaan genre tidak bisa menunjukkan perbedaan karakteristik *cluster* 1 dan *cluster* 2.



Gambar. 18. Scatterplot variabel Metascore dan Votes setiap *cluster*

Gambar 18 menunjukkan visualisasi untuk melihat pembagian *cluster* bila dilihat dengan *scatterplot* dan *density plot* variabel Metascore dan Votes. Pembagian *cluster* terlihat cukup terpisah antara *cluster* 1 dan 2 dari variabel Votes. *Cluster* 1 memiliki rata-rata jumlah votes lebih banyak dari *cluster* 2. Sedangkan bila dilihat dari nilai Metascore, kedua *cluster* memiliki rata-rata nilai Metascore yang tidak jauh berbeda.



Gambar. 19. Scatterplot variabel Runtime Gross setiap *cluster*

Gambar 19 menunjukkan visualisasi untuk melihat pembagian *cluster* setiap data dalam *scatterplot* variabel Runtime dan Gross. Dapat diketahui data *cluster* 1 tersebar dan memiliki rata-rata Runtime sedikit lebih tinggi daripada *cluster* 2 yang datanya lebih terpusat dan memiliki rata-rata Runtime lebih rendah. Density plot untuk variabel Gross juga menunjukkan rata-rata Gross untuk *cluster* 1 lebih tinggi daripada *cluster* 2.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Beberapa kesimpulan yang dapat diambil dari hasil analisis ini adalah :

a. Genre Drama menjadi genre yang paling banyak mengisi daftar 500 film teratas versi IMDb.
b. Film dengan kategori capaian rating Sangat Baik cenderung memiliki rata-rata Runtime dan Gross yang tinggi dibandingkan dengan film dengan capaian rating Baik dan Cukup Baik.
c. Steven Spielberg dan Christopher Nolan menjadi director terbaik dari jumlah film yang berhasil masuk 500 film teratas versi IMDb. Kedua director memiliki keunggulan yang berbeda dimana Steven Spielberg unggul dari jumlah film yang masuk 500 film teratas, sedangkan Christopher Nolan unggul dari rata-rata jumlah Votes yang didapat setiap filmnya.
d. Dari ketiga metode yang digunakan untuk menentukan jumlah *cluster*, didapatkan 2 sebagai nilai optimal untuk analisis *cluster* menggunakan metode *K-Means clustering*.
e. Dari analisis *K-Means clustering*, Dapat diketahui bahwa terdapat 401 film yang terkelompok di *cluster* 1, dan 99 film terkelompok menjadi *cluster* 2. *Cluster* 1 memiliki nilai median dan mean yang lebih rendah untuk semua variabel dibandingkan dengan *cluster* 2.
f. Visualisasi *bar chart* menunjukkan bahwa film dengan genre *sport* dan *musical* masuk di *cluster* 1, sementara film dengan genre lainnya tersebar ke kedua *cluster*. Diketahui perbedaan genre tidak bisa menunjukkan perbedaan karakteristik *cluster* 1 dan *cluster* 2.

B. Saran dan Rekomendasi

Saran dan rekomendasi yang dapat penulis berikan untuk penelitian selanjutnya diantaranya melakukan analisis eksplorasi data yang lebih dalam untuk mendapatkan informasi yang lebih banyak dari data yang bisa didapatkan dengan *web scraping* karena penulis belum mengeksplorasi semua variabel yang dapat di *scraping*, seperti variabel Aktor, Sinopsis, dan klasifikasi batas usia minimal film. Selain itu penelitian selanjutnya diharapkan melakukan eksplorasi metode analisis *cluster* lain yang melibatkan variabel-variabel kategorik sebagai variabel penentuan *cluster*.

Berdasarkan laman IMDb yang digunakan penulis untuk mengumpulkan data *Top 1000 IMDB US Box Office Movies* hanya terdapat 581 film yang memiliki data Metascore. Mengingat bahwa data tersebut adalah 1000 film teratas versi IMDb, penulis merekomendasikan IMDb sebagai penyedia informasi film dan acara televisi yang populer agar mengumpulkan kritik dan review film dari kritikus-kritikus terpercaya agar dapat melengkapi data Metascore pada film-film yang belum memiliki Metascore.

DAFTAR PUSTAKA

[1] M. Turland, Php | Architect's Guide to Web Scraping defined, Los Angeles, 2010.

[2] E. Prasetyo, Data Mining : Konsep dan Aplikasi Menggunakan MATLAB, Yogyakarta: Penerbit Andi, 2012.

[3] "About IMDb," IMDb, [Online]. Available: https://www.imdb.com/pressroom/?ref_=ft_pr. [Accessed 5 January 2021].

[4] R. Walpole, R. Myers, S. Myers and K. Ye, Probability & Statistics for Engineers & Scientists (9th ed.), New Jersey: Pearson Educational, Inc., 2011.

[5] C. O'Neil and R. Schutt, "Doing Data Science," O'Reilly, 2013, pp. 34-37.

[6] P. N. Tattar, S. Ramaiah and G. B. Manjunath, A course in Statistics with R, United Kingdom: John Wiley & Sons Ltd, 2016.

[7] S. Glen, "Bar Chart / Bar Graph: Examples, Excel Steps & Stacked Graphs," Statistics How To, [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/. [Accessed 8 January 2021].

[8] J. W. Tukey, "Exploratory Data Analysis: Past, Present, and Future," Princeton University, Princeton, 1993.

[9] J. NENADÁL, Měření v systémech managementu, Praha: Management Press, 2004.

[10] C. A. Janicak, Applied Statistics in Occupational Safety and Health, United Kingdom: Government Institutes, 2007.

[11] S. Nanjudan, S. Sankaran, C. R. Arjun and G. P. Anand, Identifying the Number of *Cluster*s for K-Means : A Hypersphere Density Based Approach, Chennai.

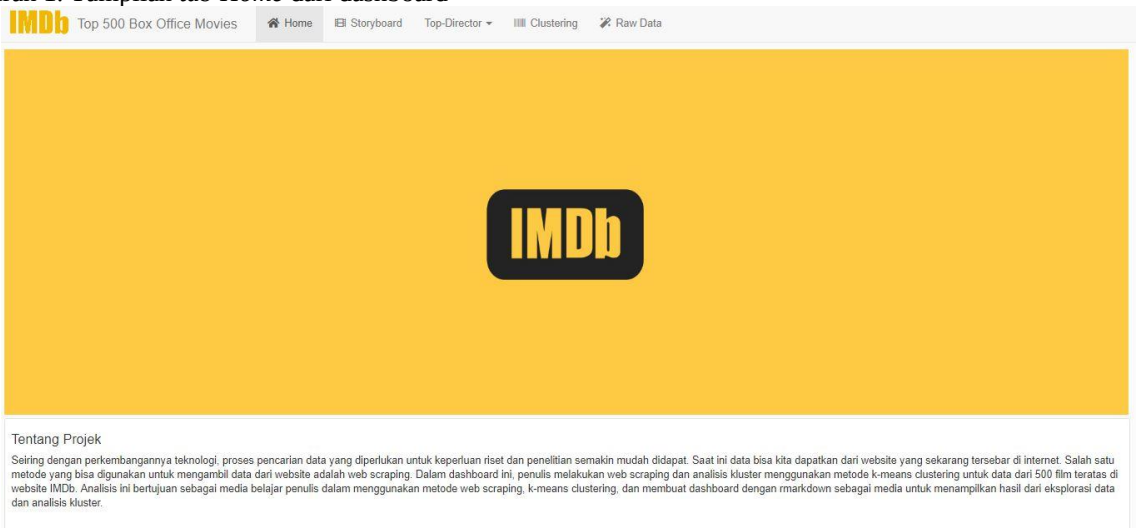
[12] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of *Cluster*s in a Data Set," *Journal of Statistical Software*, pp. 1-36, 2014.

[13] H. Giyanto, "Penerapan Algoritma *Cluster*ing K-Means, K-Medoid, Gath Geva," Tidak Terpublikasi, Yogyakarta, 2008.

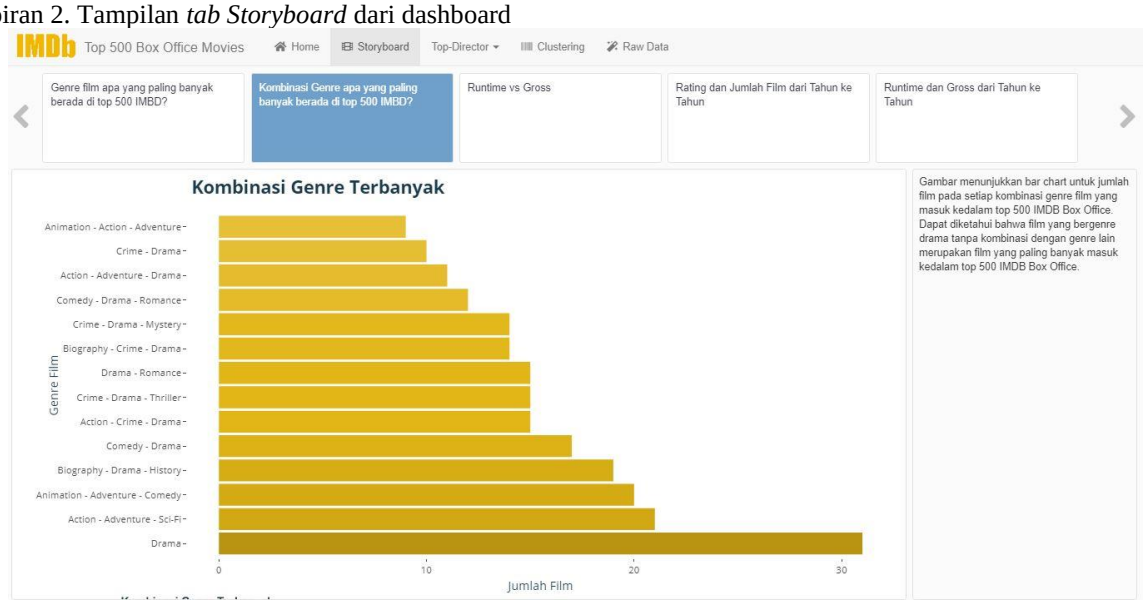
[14] J. Han and M. Kamber, Data Mining Concepts and Techniques Second Edition, San Francisco: Morgan Kauffman, 2001.

LAMPIRAN

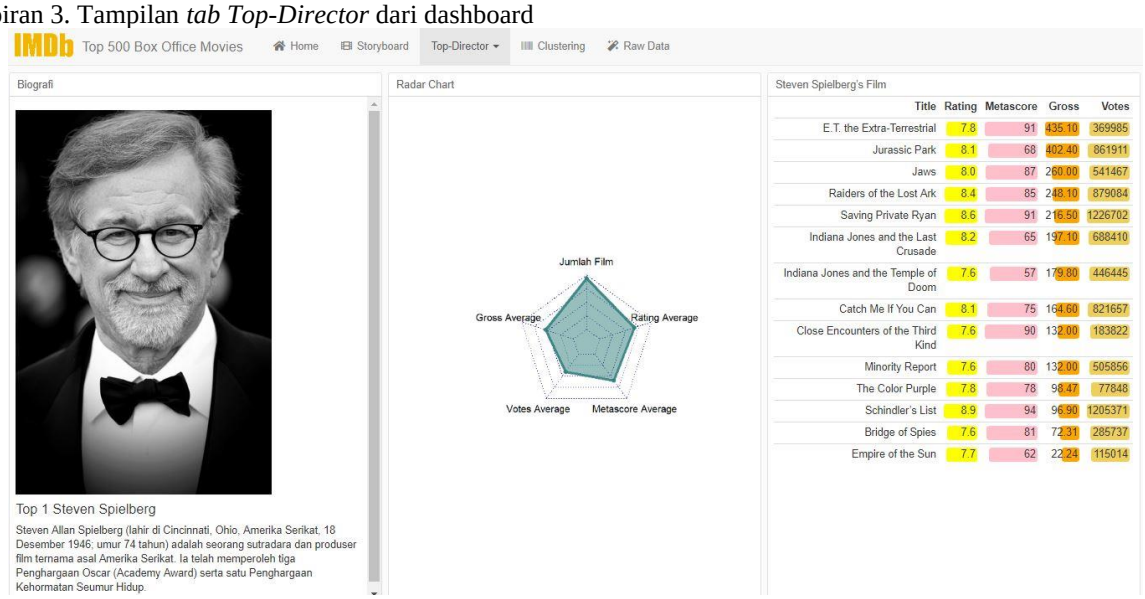
Lampiran 1. Tampilan *tab Home* dari dashboard



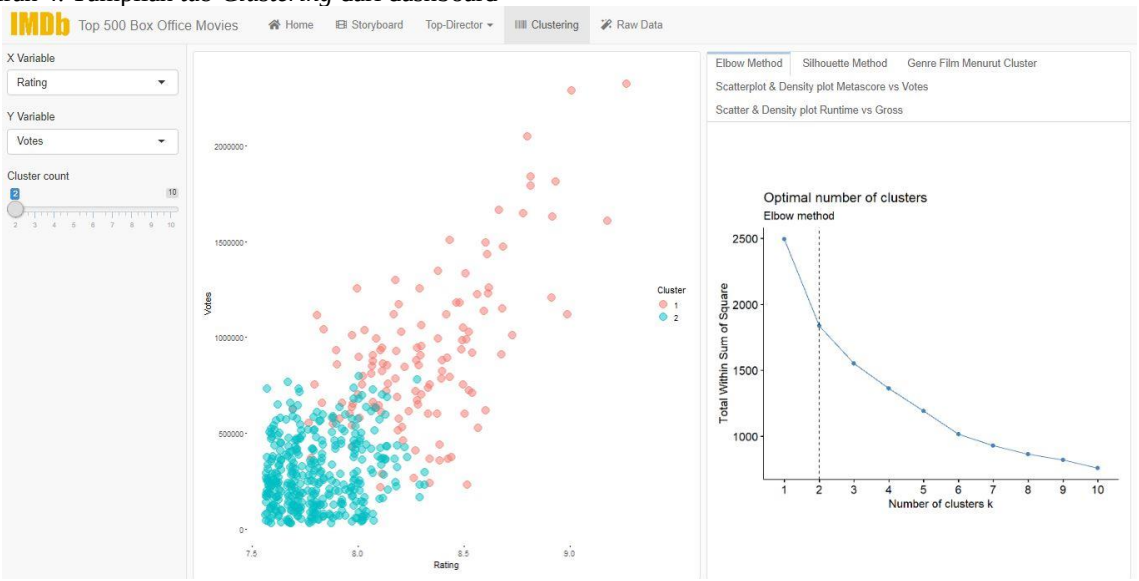
Lampiran 2. Tampilan *tab Storyboard* dari dashboard



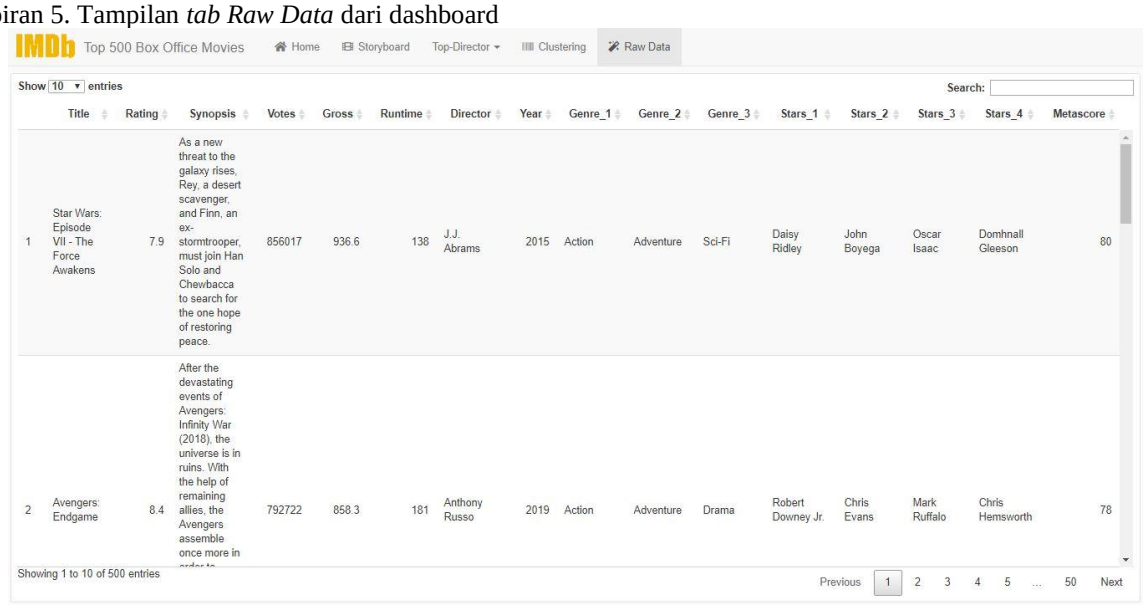
Lampiran 3. Tampilan *tab Top-Director* dari dashboard



Lampiran 4. Tampilan *tab Clustering* dari dashboard



Lampiran 5. Tampilan *tab Raw Data* dari dashboard



	Title	Rating	Synopsis	Votes	Gross	Runtime	Director	Year	Genre_1	Genre_2	Genre_3	Stars_1	Stars_2	Stars_3	Stars_4	Metascore
1	Star Wars: Episode VII - The Force Awakens	7.9	As a new threat to the galaxy rises, Rey, a desert scavenger, and Finn, an ex-stormtrooper, must join Han Solo and Chewbacca to search for the one hope of restoring peace.	856017	936.6	138	J.J. Abrams	2015	Action	Adventure	Sci-Fi	Daisy Ridley	John Boyega	Oscar Isaac	Domhnall Gleeson	80
2	Avengers: Endgame	8.4	After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in	792722	858.3	181	Anthony Russo	2019	Action	Adventure	Drama	Robert Downey Jr.	Chris Evans	Mark Ruffalo	Chris Hemsworth	78