**Penerapan dari Algoritma K-Means untuk**

**Mengelompokkan ( Clustering ) Genre Film dengan Rating Tertinggi**



**Anggota Kelompok Hitam 4PSIA :**

* **Frederick**
* **Kelven Liong**
* **Nicky Dwi Putra**
* **Ilham Noer Dwi Lony**
* **Nico Dwi Putra**

**PENDAHULUAN**

1. **Latar Belakang**

Kami melihat suatu masalah yang bisa terjadi pada UMKM kami yang bergerak pada bidang penyiaran / bioskop, yaitu Kinos Batam yang adalah sebuah mini-cinema. Sebagai UMKM, tentunya kondisi finansial dari usaha tersebut tidak setinggi seperti PT yang sudah besar seperti Cinema21, dll.

Disisi lain, saat ini Indonesia & Batam juga masih dalam situasi covid yang mengharuskan para pemilik usaha untuk lebih berhati-hati dalam menggelontorkan dana untuk ide usaha-usahanya. Saat ini, adalah saat dimana masyarakat mulai lebih membuka diri dari awal-awal covid dimana mulai menutup diri. Ini adalah situasi dan kesempatan untuk UMKM kami bergerak dan mulai menghasilkan profit.

Melihat dari perspektif pemilik usaha dalam bidang penyiaran/bioskop, tentu ada harga yang dibayar untuk mendatangkan film-film yang akan kita suguhkan kepada para pengunjung, membayar dan mengharapkan keuntungan yang diperoleh dari perilisan film tersebut serta modal usaha yang telah dikeluarkan.

Problemnya adalah terkadang pemilik usaha penyiaran memilih film-film yang akan dirilis berdasar dari kesimpulan perseorangan dibanding data, contohnya memilih film/genre film yang disukai oleh dirinya ataupun tim dalam usaha tersebut. Tidak jarang juga menebak-nebak apa yang lagi tren dan disukai oleh pelanggan. Resikonya, apabila telah mengeluarkan segala biaya yang diperlukan dan ternyata hasil tidak sesuai ekspetasi UMKMnya akan mengalami kerugian / membuang potensi keuntungan yang bisa diraih. Penyebabnya? Rata-rata adalah karena masyarakat tidak berminat pada film / genre film tersebut.

1. **Solusi**

Disinilah kami membantu UMKM kami dengan menyediakan sebuah solusi. Menurut kami dibanding merilis film / memilih genre-genre film yang ingin dirilis berdasar dari kesimpulan yang tidak didasari oleh data ataupun kesimpulan perseorangan. Lebih baik apabila keputusannya didasari oleh data yang diperoleh dari pelanggan atau masyarakat mengenai genre apa yang disukai. Dan UMKM akan merilis genre film berdasar atas jawaban-jawaban para masyarakat / pelanggan.

Perolehan data nantinya mungkin juga mengeluarkan dana. Namun tentu timbal balik yang dihasilkan lebih baik dibanding salah memilih genre film / film yang ingin dirilis yang bisa berujung pada kerugian ataupun potensi profit yang terbuang. Keuntungannya bukan terbatas pada profit / duit semata, namun juga bisa berupa engagement kepada masyarakat untuk lebih mengenal bahwa di Batam terdapat opsi lain jika ingin menonton film kesukaan mereka, yaitu UMKM kami.

Sebagai hasil dari solusi nanti akan berupa rekomendasi genre-genre film yang bisa UMKM pilih untuk keputusan dalam perilisan. Karena Scope merekomendasi suatu film terlalu kecil maka kami memilih genre film sebagai hasil rekomendasi yang akan kami berikan setelah penelitian dengan data yang diterima.

Dalam pembahasan nanti, kita akan mencontohkan sample data yang akan kita gunakan sebagai elemen pertimbangan pembuatan kesimpulan pada keputusan untuk merilis film dengan genre film yang berdasar atas data pada UMKM kami.

**METODE PENELITIAN**

**A. Pengenalan atas K-Means dan Clustering**

**K-Means Clustering** adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi ( cluster ).

Metode **K-Means Clustering** berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.

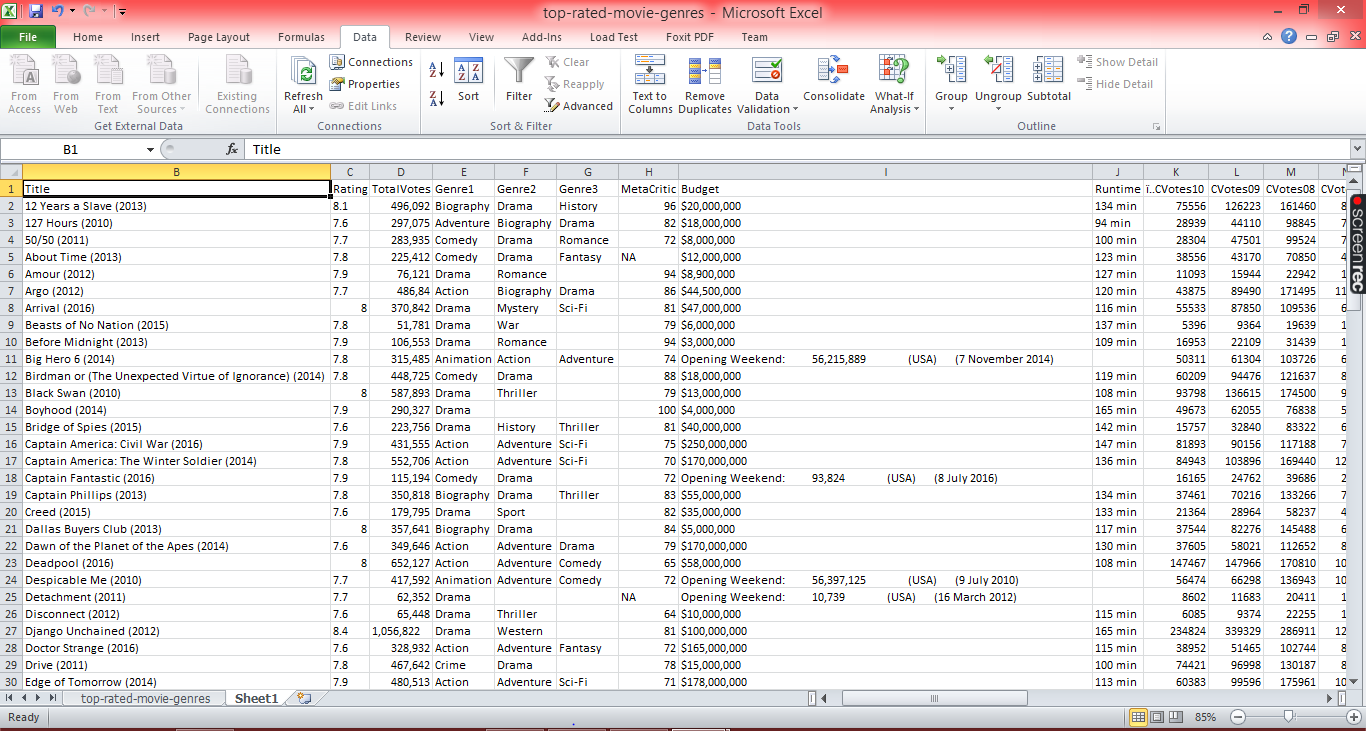
**B. Pengambilan Data**

Data yang akan diambil berupa data yang berisi tentang rating dari masyarakat terhadap genre-genre film yang disukai. Singkatnya, data yang dicari adalah data yang bisa merepresentasikan genre film apa yang sedang populer. Dalam contoh kali ini kami akan mencari data yang ingin dicari melalui website <https://data.world> , disini terdapat berbagai macam data dari berbagai belahan dunia.

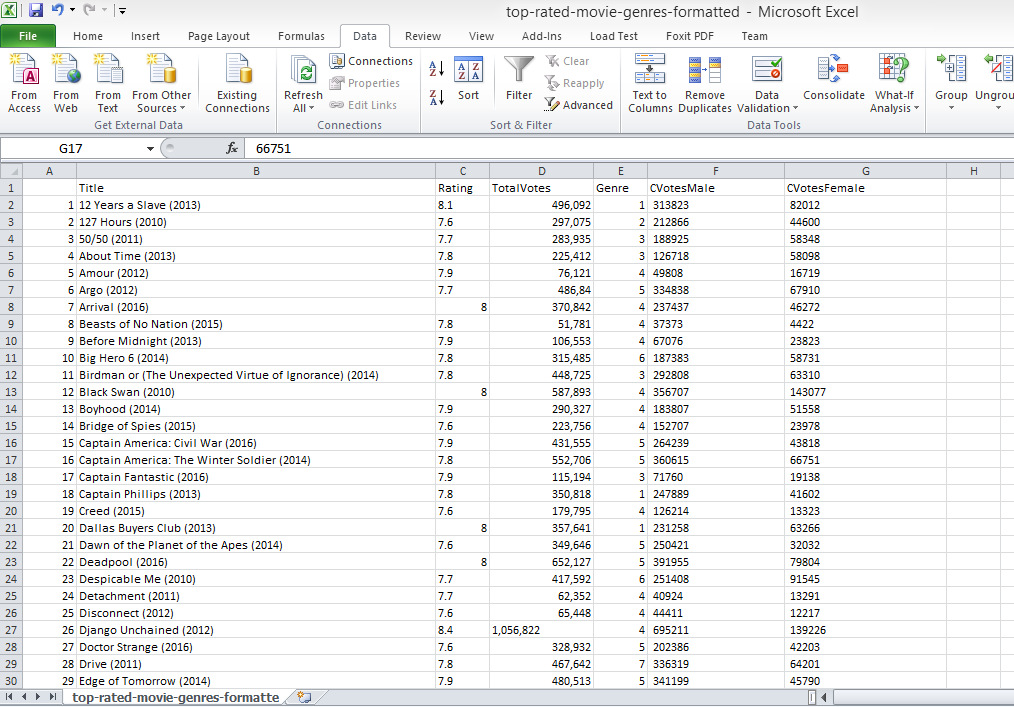
**C. Pre-processing data sebelum pengaplikasian algoritma**

Setelah menemukan data yang hendak digunakan, maka kita akan memasuki data pre-processing data. Data preprocessing adalah proses manipulasi dataset sebelum diinput kedalam model. Tujuan dari preprocessing adalah untuk membuat kesesuaian sebelum data tersebut diproses. Berikut adalah perbandingan data yang belum melalui pre-processing dan yang telah melalui pre-processing :

( Sebelum melalui pre-processing )



( Sesudah melalui pre-processing )



Dalam proses pre-processing ini, data yang didownload dan diterima dimodifikasi menjadi data yang sesuai untuk diproses nantinya. Dalam contoh ini, kami mengambil column / field yang diperlukan supaya memudahkan pembacaan data, kemudian mengubah value yang sekiranya akan sangat diperlukan dalam data processing menjadi angka untuk kemudahan dalam pengaplikasian dengan algoritma yang dipilih nantinya.

Value yang dimodifikasi adalah value dari genre itu sendiri, jadi setelah memodifikasi kami mendapati terdapat 8 macam genre dalam data ini. Berikut adalah data setelah termodifikasi tersebut, dimana setiap angka yang tertera mewakili value dari genrenya :

1 - Biography

2 - Adventure

3 - Comedy

4 - Drama

5 - Action

6 - Animation

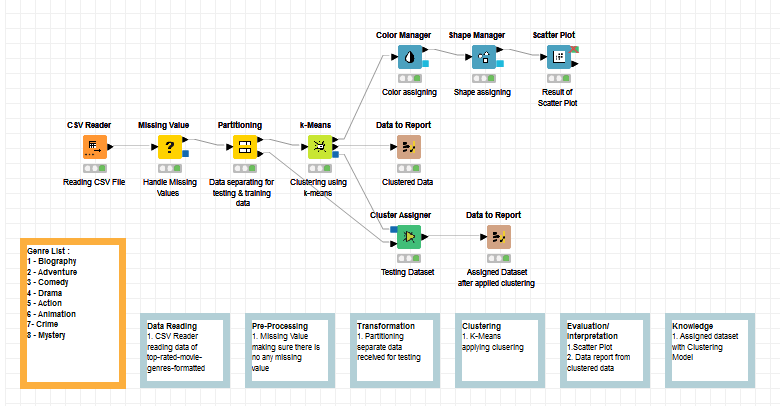
7- Crime

8 - Mystery

**HASIL DAN PEMBAHASAN**

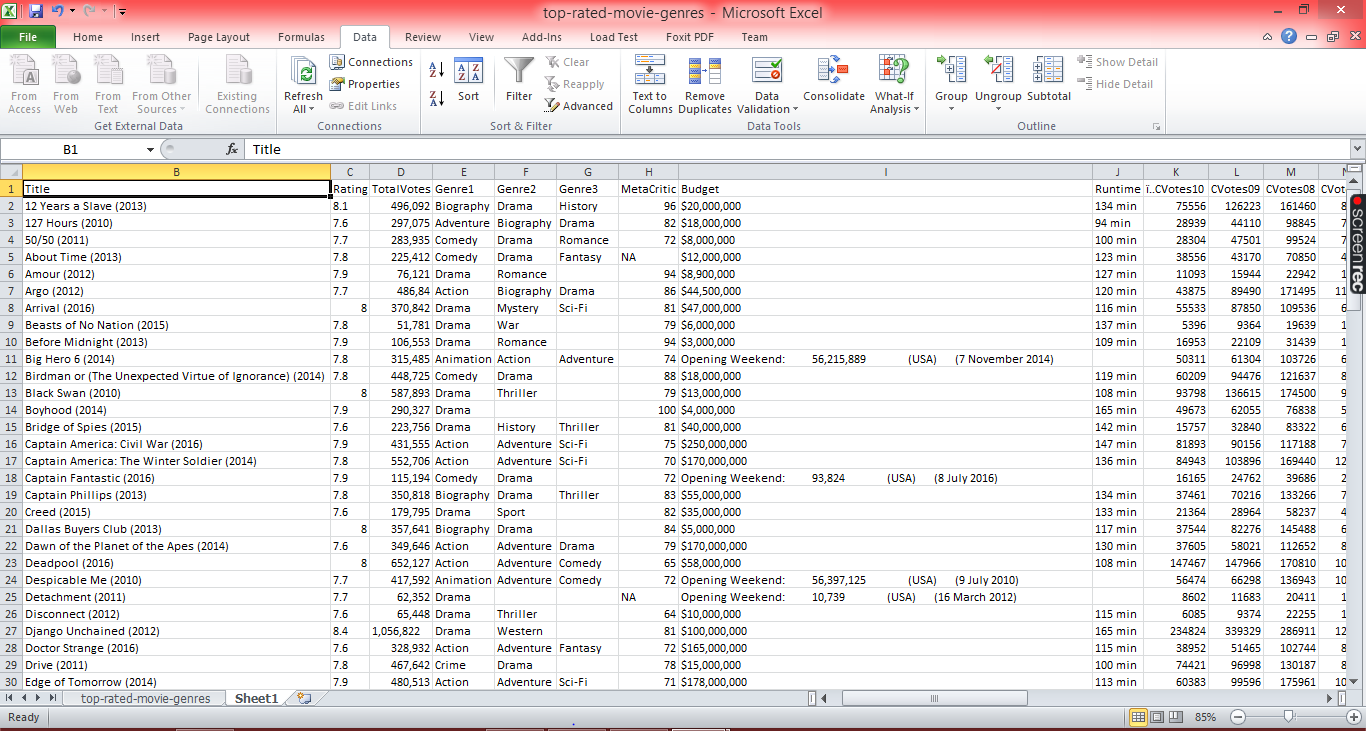
Dalam penelitian ini, genre & rating akan menjadi objek dari penelitian. Dengan data yang diterima berisi genre dengan rating, judul, total votes, dan juga total votes dari pria dan wanita, kami akan menggunakan aplikasi KNIME dengan algoritma K-Means Clustering.

Berikut adalah gambar yang merepresentasikan langkah-langkah dari penelitian yang dilakukan dalam aplikasi KNIME dengan algoritma K-Means Clustering.



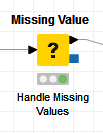
**A. Pemilihan Data**

Data yang kami pilih adalah data yang kami terima dari <https://data.world/saipranav/top-rated-english-movies-of-this-decade-from-imdb/workspace/file?filename=IMDB.csv> , dimana berisi data top rated movies dari imdb ( internet movie database ). Berikut adalah snapshot dari datasetnya :



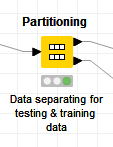
**B. Pre-Processing Data**

Bila diatas adalah pre-processing data sebelum pengaplikasian algoritma maka ini adalah pre-processing dalam pengaplikasian algoritma atau didalam aplikasi KNIME. Dalam aplikasi KNIME kami melakukan pre-processing data dengan Missing Value. Dimana dengan ini akan ada cleaning data untuk merapikan data dari value yang kosong



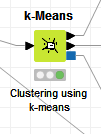
**C. Dataset Transformation**

Pada proses penelitian tahap ini, kita memisah dataset menjadi rasio 9:1 menggunakan node partitioning. Jadi dari dataset utama akan ada 90% data sebagai data training dan 10% sebagai data testing dimana data training akan dikirim ke node berikutnya ke dalam algoritma clustering dari K-Means, dan data testing akan digunakan dalam tahap percobaan/testing.

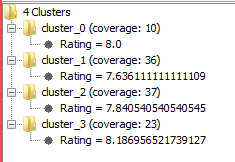
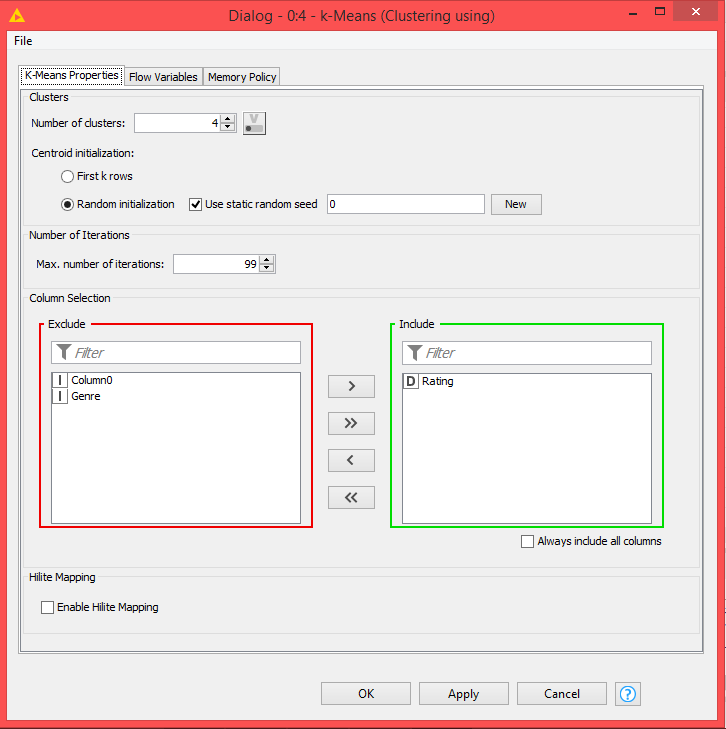


**D. Clustering dengan algoritma K-Means**

Sekarang dari dataset yang telah diterima, akan kita integrasikan dengan algoritma Clustering dari K-Means. Data-data yang akan dikumpulkan bisa kita setting dan kita tentukan berapa cluster yang ingin kita bikin dengan dataset yang diterima.



Dalam setting kali ini kami akan membentuk 4 cluster dan hanya mengambil rating. Sehingga data-data yang masuk akan kami ambil data rating dari 4 cluster. Berikut adalah hasilnya :



Dari sini, kita bisa melihat bahwa terdapat 4 cluster hasil dari setting kita, yaitu :

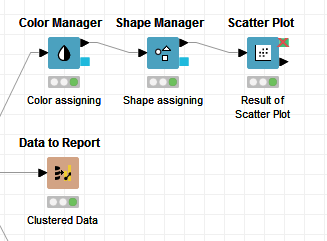
1. cluster\_0, dengan rating 8.0  
2. cluster\_1, dengan rating 7.6361…  
3. cluster\_2, dengan rating 7.8405…

4. cluster\_3, dengan rating 8.1869…

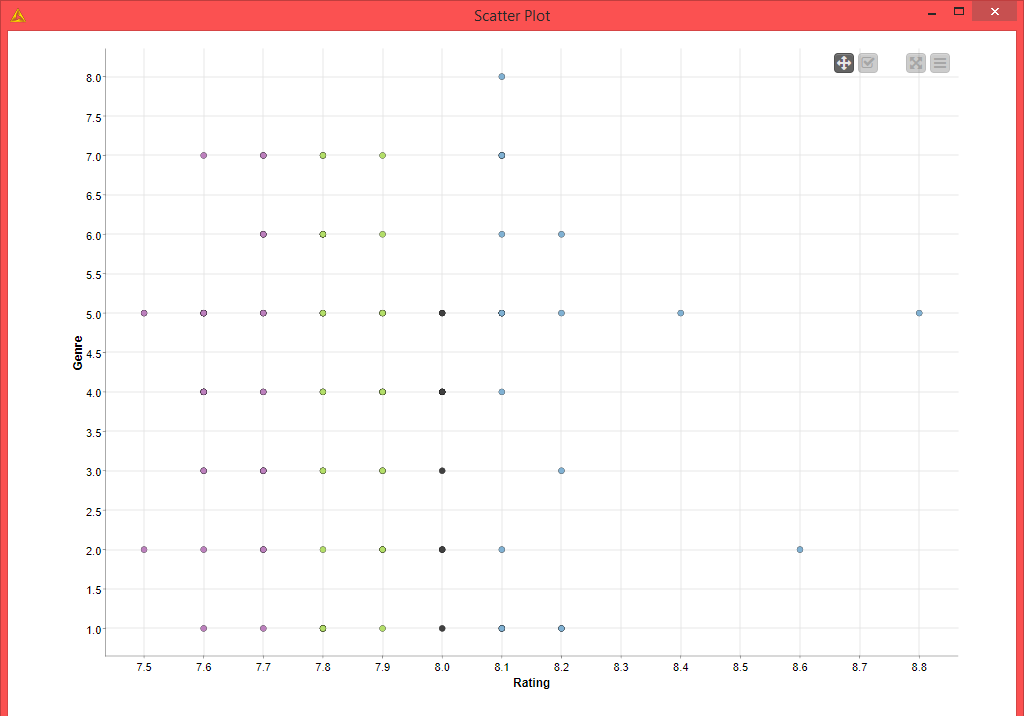
Kita akan menggunakan ini sebagai referensi dalam evaluasi dan interpretasi tahap selanjutnya.

**E. Evaluasi dan Interpretasi**

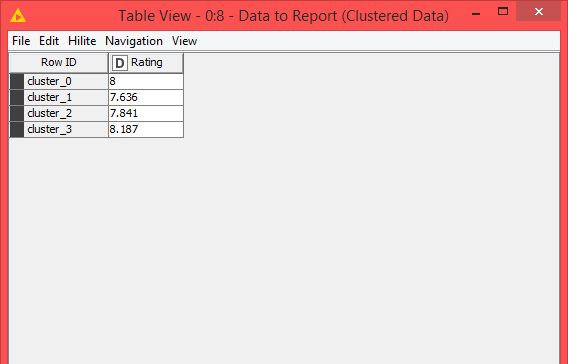
Pada tahap ini, hasil clustering akan didistribusikan ke 2 node. Node pertama adalah Scatter plot, dimana sebelum ke scatter plot kita perlu setting warna dan bentuk yang akan masuk ke scatter plot nantinya.



Ini adalah hasil dari scatter plot berdasar atas rating ( X axis ) & genre ( Y axis ).



Dan kemudian akan didistribusikan ke node data-to-report :



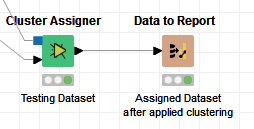
Dari gambar diatas kita bisa melihat bahwa hasil cluster kita bisa kita definisikan dalam arti :

1. cluster\_3 = Rating Sangat Baik  
2. cluster\_0 = Rating Baik  
3. cluster\_2 = Rating Standar

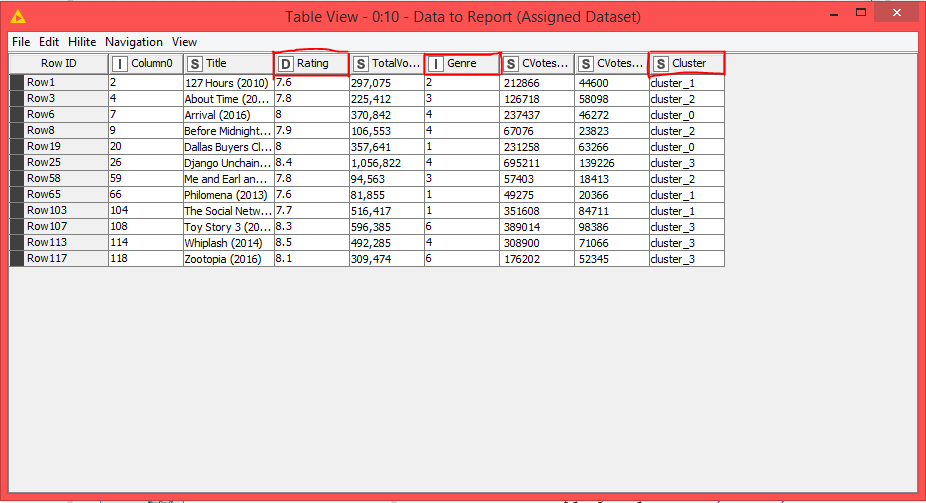
4. cluster\_1 = Rating Cukup

**F. Pengetahuan**

Sampailah kita pada tahap untuk mengetes testing data yang telah kita pisahkan dengan model clustering yang kita buat.



Data testing yang dites ini memiliki column/field yang sama dengan training data namun dengan value yang berbeda, singkat cerita untuk kedepannya apabila menerima data baru dari pelanggan/calon pengunjung bisa kita terima dan rubah data mengikuti format contoh data yang kita miliki. Berikut adalah hasil dari pengaplikasian clustering model yang telah dibuat dengan testing data.



Pada tahap **Clustering dengan algoritma K-Means**, kita bisa melihat bahwa setiap cluster memiliki valuenya masing-masing. Dari gambar diatas, kita dapat membandingkan dan melihat field rating & cluster, contohnya 7.6 masuk ke cluster\_1 , 8.0 masuk ke cluster\_0 dan 8.5 dimana adalah rating tertinggi masuk ke cluster\_3 yang juga adalah cluster dengan value tertinggi.

Melihat hasil dari gambar diatas yang kita bandingkan, analisis dan perhatikan, dapat kita simpulkan bahwa clustering model yang diintegrasikan kepada data untuk testing berhasil.

**KESIMPULAN**

Dari hasil penelitian yang telah dijalankan pada dataset rating dari masyarakat terhadap genre-genre film yang disukai. Dapat kita simpulkan bahwa :

1. Genre **Drama** dengan value 4 pada dataset yang telah melalui pre-processing mendapat rating tertinggi. Sehingga dengan itu UMKM Kinos Batam dapat lebih mempertimbangkan untuk merilis film-film dengan genre Drama dalam perencanaan mendatang.
2. Genre **Animation** dengan value 6 pada dataset yang telah melalui pre-processing menduduki posisi ke-2 sebagai genre dengan rating tertinggi. Dengan demikian, UMKM kinos-batam dapat menjadikan genre Animation sebagai alternatif / pilihan genre lain dalam perencanaan merilis film-film pada masa mendatang.
3. Dengan menerima masukan mengenai genre-genre film dengan rating tertinggi, maka kemungkinan dari UMKM Kinos Batam untuk mendatangkan genre dengan hasil profibilitasnya kecil bahkan merugi bisa lebih rendah, dan kemungkinan terbaiknya bisa memaksimalkan profibilitas.
4. Dengan model clustering yang sudah berhasil, kedepannya UMKM Kinos Batam dapat memanfaatkan ini sebagai wadah untuk menampung hasil suara dari para pelanggan maupun calon pelanggan mengenai genre apa yang mereka sukai / bagaimana rating mereka terhadap genre yang menjadi calon-calon pilihan dari UMKM Kinos Batam dan bisa menerima hasil mengenai genre apa yang menduduki posisi-posisi dengan rating tertinggi.

**DAFTAR PUSTAKA**

* <https://informatikalogi.com/algoritma-k-means-clustering/>
* <https://skillplus.web.id/data-preprocessing/#:~:text=Lesson%20data%20preprocessing%20akan%20memperkenalkan,dalam%20membuat%20model%20machine%20learning.&text=Data%20preprocessing%20adalah%20proses%20manipulasi,kompatibel%20dengan%20library%20yang%20digunakan>.
* <https://data.world/saipranav/top-rated-english-movies-of-this-decade-from-imdb/workspace/file?filename=IMDB.csv>