

Analisa Nilai K terhadap Pengelompokkan Penjualan Tiket pada K-Means dengan Metode Elbow dan Silhouette

Elbow and Silhouette Methods for K Value Analysis of Ticket Sales Grouping on K-Means

¹Adityo Putro Wicaksono*, ²Stephanus Widjaja, ³Mohamad Fajarianditya Nugroho, ⁴Christina
⁵Priscilla Putri

^{1,3,4,5}Teknik Informatika, Teknik dan Informatika, Universitas AKI

²Teknik Informatika, STMIK AKI Pati

Jalan Imam Bonjol No. 15-17, Semarang, Jawa Tengah, Indonesia

*e-mail: adityoputrowicaksono@gmail.com

(received: 16 Mei 2023, revised: 12 September 2023, accepted: 19 September 2023)

Abstrak

Persaingan bisnis yang semakin ketat memaksa orang untuk mengembangkan bisnis untuk bersaing dalam menjual dengan mencari lead untuk meningkatkan pendapatan dari penjualan bisnis. Perkembangan era digital mempengaruhi penggunaan internet untuk proses pencarian informasi dan pembelian secara online, seperti pembelian tiket dan produk digital. Film merupakan salah satu industri di dunia hiburan yang membawa suasana baru bagi masyarakat untuk menonton film yang bisa ditonton dimana saja dan juga online maupun offline. Data penjualan tiket yang sedang diproses di-cluster menggunakan K-Means Proses clustering dan evaluasi untuk menentukan nilai K terbaik pada proses clustering menggunakan *Metode Elbow* dan *Silhouette Score*. Pengelompokan diperoleh apabila hasil pengelompokan tidak berbeda nyata satu saat proses pengelompokan dilakukan pada "atribut *ticket_sold*" dan "*ticket_price*". Berdasarkan hasil *Metode Elbow*, diperoleh hasil bahwa nilai $k=3$ sampai dengan nilai $k=5$ merupakan nilai optimal untuk melakukan proses pengelompokan data untuk kasus penjualan tiket film. Sedangkan hasil pembagian cluster yang baik berada pada nilai $k=5$ dengan nilai 0,60 dengan menggunakan *Silhouette Score*.

Kata kunci: Clustering, k-Means, Elbow, Silhouette, Tiket.

Abstract

Increasingly fierce business competition forces people to develop businesses to compete in selling by looking for leads to increase revenue from business sales. The development of the digital era affects the use of the internet for the process of searching for information and online purchases, such as purchasing tickets and digital products. Film is one of the industries in the entertainment world that brings a new atmosphere for people to watch films that can be watched anywhere and also online and offline. Ticket sales data that is being processed is clustered using K-Means. The clustering process and evaluation to determine the best K value in the clustering process uses the Elbow and Silhouette Score Methods. Grouping is obtained if the grouping results are not significantly different when the grouping process is carried out on the "ticket_sold" and "ticket_price" attributes. Based on the results of the Elbow Method, the result is that the value of $k = 3$ to the value of $k = 5$ is the optimal value for carrying out the data grouping process for the case of selling movie tickets. While the results of a good cluster distribution are at a value of $k = 5$ with a value of 0.60 using the Silhouette Score..

Keywords: Clustering, k-Means, Elbow, Silhouette, Ticket.

1 Pendahuluan

Persaingan bisnis yang semakin kompetitif membuat masyarakat diharuskan untuk mengembangkan usaha demi keberlangsungan hidup dalam persaingan penjualan dengan cara menemukan petunjuk untuk peningkatan pendapatan dalam penjualan perusahaan. Selain itu juga,

<http://sistemasi.ftik.unisi.ac.id>

perkembangan era digital berdampak pada penggunaan internet dalam pencarian informasi dan proses pembelian berbasis online seperti pembelian tiket dan produk digital [1][2]. Film merupakan salah satu industri di dunia hiburan yang memberikan suasana baru bagi masyarakat untuk menonton film yang bisa diakses dimanapun dan juga bisa ditonton secara *online* maupun *offline*. Film memiliki berbagai jenis judul, genre, dan juga peminat dari masing-masing penonton. Dalam penjualan tiket film, penentuan harga menjadi salah satu tolak ukur di dalam pemasaran yang paling berpengaruh terhadap peningkatan penjualan [3]. Selain itu juga, pentingnya proses pengambilan keputusan di dalam pembelian tiket guna untuk segmentasi pelanggan pun juga diperlukan agar lebih mudah di dalam mengidentifikasi pelanggan berdasarkan kebiasaan pelanggan terhadap film yang disukainya [4][5].

Seiring dengan berjalannya waktu, di dalam melakukan proses segmentasi pelanggan diperlukan sebuah cara untuk melakukan proses segmentasi pelanggan berdasarkan data penjualan tiket film yang terdiri dari berbagai kriteria dan record data yang banyak dengan baik, tepat, dan dapat digunakan untuk pengambilan keputusan. Pengambilan keputusan ini akan digunakan untuk menentukan strategi pemasaran berdasarkan kebiasaan pelanggan terhadap beberapa film yang sudah dibeli dan ditonton dari segi harga, jumlah jam tayang, total penjualan dan pembelian tiket, dan beberapa atribut lainnya. Selain itu, proses segmentasi akan dilakukan dengan mengelompokkan pelanggan berdasarkan dari atribut yang ditentukan diatas.

Data penjualan tiket yang akan diolah saat ini, akan dilakukan proses pengelompokkan dengan menggunakan K-Means dikarenakan metode ini dapat menentukan proses pengelompokkan berdasarkan atribut atau fitur dari objek ke dalam K kluster [5][6]. Selain menentukan proses pengelompokkan, akan dilakukan proses evaluasi untuk mencari nilai K terbaik di dalam proses pengelompokkan klusterisasi dengan menggunakan Metode *Elbow* dan *Silhouette Coefficient*, yang dimana metode *Elbow* digunakan untuk menentukan pola kluster terbaik dari dilai K yang ditentukan sebanyak N kluster [4] dan *Silhouette Coefficient* untuk melihat seberapa baik kualitas dari hasil pengelompokkan atau kluster dan seberapa efektif dari data yang sudah dikelompokkan dalam kluster yang sudah dibuat [7].

Dengan menerapkan proses klustering dengan menggunakan k-means untuk penentuan segmentasi penonton film, diharapkan bisa menjadi bahan pertimbangan di dalam menentukan strategi bisnis dan pengambilan keputusan untuk penjualan tiket film kedepan dan dapat diterapkan menjadi pengembangan sistem rekomendasi film yang sudah rilis.

2 Tinjauan Literatur

Penelitian sebelumnya terkait penjualan produk digital menggunakan metode clustering menggunakan algoritma K-Means menggunakan euclidean distance untuk menentukan jarak antar cluster terhadap centroid dan ditentukan menjadi 3 pembagian cluster dalam penelitian Astuti dkk [1]. Pengintegrasian metode K-Means untuk mencari nilai K terbaik dengan menggunakan Elbow Method dilakukan untuk proses identifikasi kluster profil pelanggan dengan penentuan 500 data kluster sehingga ditemukan untuk nilai K terbaik menurut Syakur dkk yaitu nilai $K=3$ [4]. Proses segmentasi pelanggan juga dilakukan oleh Nisa dkk dengan menggunakan K-Means untuk profiling pelanggan. Untuk pengujian data training dan testing dibagi menjadi 3 simulasi dengan pembagian data 80%-20%, 70%-30%, dan 50%-50% [8].

Pada penelitian yang dilakukan oleh Billah dkk, untuk pembangunan sistem rekomendasi film dengan menggunakan K-Means dengan Collaboration Filtering dan PCA dengan dataset yang berbeda diperoleh hasil rekomendasi dengan menggunakan Mean Reciprocal Rank (MRR) memiliki hasil rekomendasi dari cluster yang cukup tepat, yaitu sekitar 0,44 sampai dengan 0,56[9], [10]. Selain itu, pengelompokkan data untuk identifikasi kemiskinan di daerah Jawa Barat yang dilakukan oleh Nabilla dkk menggunakan pendekatan berbasis Data Mining dengan CRISP-DM sehingga dari hasil yang diperoleh menunjukkan bahwa hasil cluster yang terbaik dengan menggunakan silhouette method ditunjukkan dengan nilai $K=2$ dengan nilai silhouette score = 0,576[11].

Berdasarkan hasil penelitian dari beberapa peneliti sebelumnya, belum ada yang membahas tentang efektifitas nilai K untuk penentuan klustering sehingga belum ditemukan hasil yang baik dan efektif untuk pemanfaatan nilai K menggunakan silhouette dan elbow method. Sehingga pada penelitian ini dilakukan dengan membandingkan nilai K yang paling baik untuk proses klustering dalam penentuan harga penjualan tiket film.

K-Means Clustering merupakan algoritma clustering yang mengelompokkan data ke dalam kelompok-kelompok dengan karakteristik yang mirip. Algoritma ini menggunakan metode pengukuran jarak sebagai alat ukur kedekatan dua buah kumpulan data. Berikut adalah langkah-langkah dari algoritma Clustering K-Means:

1. Tentukan nilai k , misal jumlah cluster yang akan dibentuk.
2. Acak k centroid dari data yang ada. Pusat awal data dapat dipilih secara acak atau dengan metode yang lebih canggih.
3. Kelompokkan semua data ke dalam kelompok centroid terdekat menggunakan jarak Euclidean.
4. Hitung kembali pusat massa setiap kelompok dengan rerata data dalam kelompok tersebut.
5. Ulangi langkah 3 dan 4 hingga centroid tidak berubah atau batas iterasi yang diperlukan tercapai.

Ketika algoritma clustering sudah siap, hasil cluster yang terbentuk dapat dievaluasi menggunakan metrik evaluasi seperti Sum of Squared Errors (SSE), Silhouette Score, atau Davies-Bouldin Index. Pada evaluasi algoritma clustering, memiliki beberapa definisi evaluasi, seperti SSE untuk mengukur jumlah kesalahan kuadrat semua data dalam grup dan pusat massa. Silhouette Score dan Davies-Bouldin Index digunakan untuk mengukur kualitas klaster berdasarkan data dalam grup dan antar grup [12].

Metode *Elbow* digunakan dalam proses pengujian jumlah k pada model klustering dan memiliki peranan yang penting. Salah satu kelemahan pada K-Means adalah dalam penentuan jumlah k terbaik dari setiap n percobaan yang dilakukan. Dengan urgensi tersebut, maka Metode *Elbow* menjadi solusi pemodelan K-Means di dalam penentuan nilai k [4]. Persamaan untuk Metode *Elbow* dapat dituliskan pada persamaan 1 berikut.

$$d = \sum(x_i - t_x) + (y_i - t_y) \quad (1)$$

Dengan keterangan:

d = nilai *Distortion*

x_i = Cluster (x) pada perulangan ke (i)

t_x = Titik tengah *cluster* (x)

y_i = Cluster (y) pada perulangan ke (i)

t_y = Titik tengah *cluster* (y)

Silhouette score berperan dalam pengukuran hasil cluster. Metode ini sebagai alternatif pemilihan k terbaik untuk pemodelan K-Means, sehingga model yang dibuat berdasarkan nilai *Silhouette Score* dengan nilai tinggi merepresentasikan struktur data yang telah dikelompokkan [13]. Berikut persamaan dari *Silhouette Score* yang dituliskan pada persamaan 2.

$$s(a) = y - x \times \max(x, y) \quad (2)$$

Dengan keterangan:

$s(a)$ = Nilai *Silhouette Score*

x = Rata-rata nilai *intra cluster distance*

y = Rata-rata nilai *inter cluster distance*

CRISP-DM merupakan model proses untuk melihat korelasi atau relasi dari berbagai tugas Data Mining berdasarkan tujuan, latar belakang dan ketertarikan pada pengguna dan terutama berdasarkan data. Sebuah proyek Data Mining memiliki enam tahapan siklus hidup. Dan setiap fase pada CRISP-DM bersifat tidak tetap. Bergerak bolak-balik di antara fase-fase tersebut selalu diperlukan. Fase mana atau tugas spesifik mana dari suatu fase yang akan diselesaikan selanjutnya ditentukan oleh hasil dari setiap fase[14]. Pada gambar 1, panah mewakili ketergantungan antar fase yang paling penting dan sering terjadi. Lingkaran luar mewakili sifat siklus dari data mining. Setelah solusi diimplementasikan, penggalian data tidak berakhir. Pelajaran yang didapat selama proses dan dari solusi yang diterapkan dapat memicu pertanyaan bisnis baru yang lebih terarah. Teknik Data Mining selanjutnya akan mendapatkan keuntungan dari pengalaman sebelumnya.

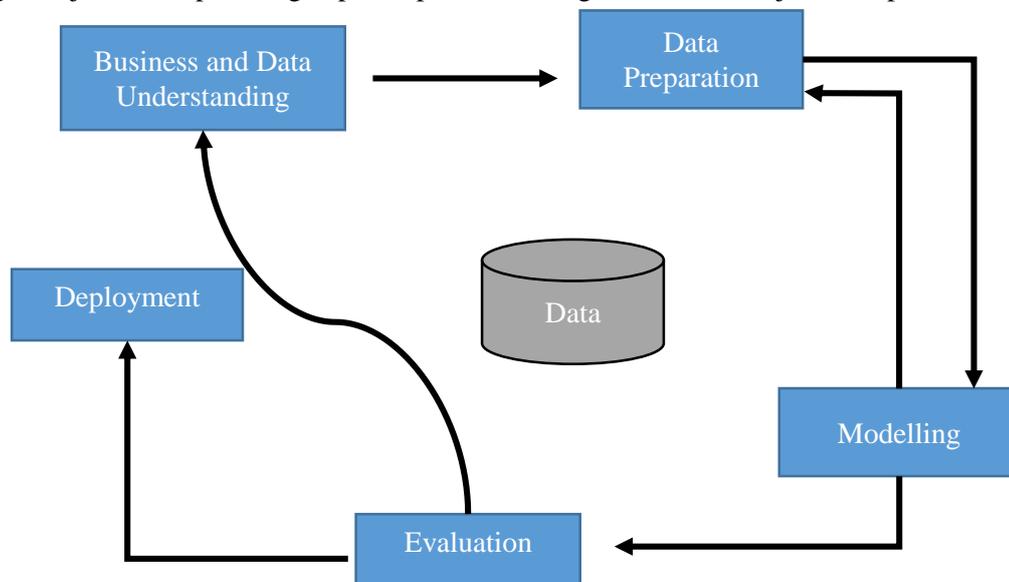
Tahapan-tahapan pada model proses CRISP-DM dibagi menjadi 6 tahapan[14]:

1. **Business Understanding**
Fase pertama ini berfokus pada pemahaman tujuan dan kebutuhan proyek dari sudut pandang bisnis, kemudian mengubah pengetahuan ini menjadi definisi masalah data mining dan rencana awal untuk memenuhi tujuan tersebut.
2. **Data Understanding**
Fase Data Understanding dimulai dengan pengumpulan data dan dilanjutkan dengan tindakan untuk mengenal data, menemukan masalah kualitas data, mengungkap wawasan pertama ke dalam data, atau mendeteksi subset yang menarik untuk membuat hipotesis untuk informasi yang tersembunyi.
3. **Data Preparation**
Fase Data Preparation mencakup semua operasi yang menghasilkan kumpulan data akhir (data yang akan dipasok ke dalam alat pemodelan) yang dibangun dari data mentah awal. Tugas persiapan data kemungkinan besar akan dilakukan beberapa kali dan tanpa urutan tertentu. Pemilihan tabel, catatan, dan atribut, serta transformasi dan pembersihan data untuk alat bantu pemodelan, merupakan tugas-tugas tersebut.
4. **Modeling**
Berbagai teknik pemodelan dipilih dan digunakan dalam fase ini, dan parameternya dikalibrasi ke nilai yang ideal. Biasanya, ada banyak strategi untuk jenis masalah penggalian data yang sama. Beberapa pendekatan memiliki persyaratan format data khusus. Akibatnya, kembali ke proses persiapan data sering kali diperlukan.
5. **Evaluation**
Pada titik ini dalam proyek, sudah ada pembuatan model (atau beberapa model) yang tampaknya berkualitas tinggi berdasarkan analisis data. Sebelum melanjutkan dengan penerapan akhir model, sangat penting untuk memeriksa model secara menyeluruh dan meninjau prosedur yang digunakan untuk mengembangkan model untuk memastikan bahwa model tersebut memenuhi tujuan bisnis. Salah satu tujuan penting adalah untuk mengidentifikasi apakah ada masalah bisnis yang penting yang terlewatkan. Keputusan tentang bagaimana menggunakan hasil penggalian data harus dibuat pada akhir langkah ini.
6. **Deployment**
Dalam banyak kasus, pembuatan model bukanlah akhir dari proyek. Meskipun tujuan model adalah untuk memperluas pemahaman data, pengetahuan yang diperoleh harus diatur dan disajikan dengan cara yang dapat digunakan oleh user. Hal ini sering kali melibatkan penggabungan model *live* ke dalam proses pengambilan keputusan organisasi, seperti penyesuaian halaman Web secara real-time atau penilaian database pemasaran yang berulang-ulang. Fase Deployment, tergantung pada tujuannya, mungkin semudah membuat laporan atau seruit memasang proses penggalian data yang berulang-ulang di seluruh perusahaan. Dalam banyak situasi, pelanggan, bukan analis data, yang melakukan proses penerapan. Bahkan jika hasil dari analisa tidak bisa mendukung proses pemahaman data, maka sangat penting bagi user untuk memahami langkah yang akan diambil untuk penggunaan model yang tepat.

Penelitian yang dibuat bertujuan untuk mengetahui nilai pengelompokan (K) terbaik di dalam pengelompokan penjualan tiket film berdasarkan beberapa atribut. Penggunaan metode K-Means memiliki kekurangan di dalam penentuan nilai K untuk pembagi kluster, sehingga dengan menggunakan *Elbow Method* dan *Silhouette Score* dapat membantu menentukan nilai K untuk pembagi cluster terbaik pada kasus ini [15]. Ketika penerapan penelitian ini berhasil, maka dapat digunakan untuk bahan pertimbangan di dalam menentukan strategi bisnis dan pengambilan keputusan untuk penjualan tiket film kedepan dan dapat diterapkan menjadi pengembangan sistem rekomendasi film yang sudah rilis.

3 Metode Penelitian

Metode penelitian untuk melakukan analisa nilai K terhadap pengelompokkan penjualan tiket menggunakan CRISP-DM. CRISP-DM merupakan salah satu proses model yang memberikan gambaran ringkas tentang pelaksanaan proses data mining maupun data science[14]. Pada penelitian ini, terbagi menjadi 5 tahapan dengan penerapan metodologi CRISP-DM dijabarkan pada Gambar 1.



Gambar 1. Tahapan Penelitian CRISP-DM

Fase-fase Penelitian dengan CRISP-DM:

1. Business and Data Understanding
Pada fase ini, dilakukan proses analisa masalah pada penjualan tiket film berdasarkan dan pemahaman data penjualan tiket film dengan melihat data dan mengidentifikasi dataset penjualan tiket *cinemaTicket-Ref.csv* dengan total data sebesar 142.524 row data dengan 14 atribut.
2. Data Preparation
Data yang sudah diidentifikasi di awal pada fase Business and Data Understanding, dilakukan proses pembersihan data dan modifikasi data diantaranya penghilangan data kotor (noisy data), penanganan missing value, reduksi outlier, dan mencari korelasi antar atribut pada dataset penjualan tiket *cinemaTicket-Ref.csv*.
3. Modelling
Setelah dilakukan persiapan data, langkah berikutnya menerapkan metode pengelompokkan data dari dataset penjualan tiket. Metode yang digunakan menggunakan K-Means Clustering dengan percobaan nilai k untuk pengelompokkan dari 3 sampai dengan 5 kelompok. Selain itu, proses pengelompokkan dengan model K-Means ini berdasarkan dari korelasi antar atribut yang sudah dilakukan pada tahapan data preparation.
4. Evaluation
Evaluasi dari hasil modeling akan dilakukan guna melihat hasil pengelompokkan yang paling baik dengan 2 buah teknik evaluasi. Teknik evaluasi yang digunakan yaitu dengan Metode Elbow dan Silhouette Score untuk melihat pembagian nilai K yang terbaik pada pengelompokkan penjualan tiket film.
5. Deployment
Setelah dilakukan proses evaluasi, maka akan dilakukan proses deployment berupa visualisasi dari hasil pengelompokkan dengan nilai K = 3 sampai dengan K=5.

4 Hasil dan Pembahasan

4.1. Hasil Data Understanding pada Dataset Penjualan Tiket Film

Dalam pengolahan data penjualan tiket film, dilakukan proses identifikasi data dan analisis data menggunakan *Google Colab* dengan bahasa pemrograman *Python*. Dikarenakan dataset yang dipakai membutuhkan proses pengolahan data yang tidak memakan waktu, maka perlu dilakukan proses analisa data menggunakan coding. Pada hasil pengolahan dataset penjualan tiket film yang ditunjukkan pada gambar 2 dan 3, diperoleh total data sebanyak 142.524 row data dengan 14 atribut selama satu tahun di tahun 2018.

	film_code	cinema_code	total_sales	tickets_sold	tickets_out	show_time	occu_perc	ticket_price	ticket_use	capacity	date	month	quarter	day
0	1492	304	3900000	26	0	4	4.26	150000.0	26	610.328638	2018-05-05	5	2	5
1	1492	352	3360000	42	0	5	8.08	80000.0	42	519.801980	2018-05-05	5	2	5
2	1492	489	2560000	32	0	4	20.00	80000.0	32	160.000000	2018-05-05	5	2	5
3	1492	429	1200000	12	0	1	11.01	100000.0	12	108.991826	2018-05-05	5	2	5
4	1492	524	1200000	15	0	3	16.67	80000.0	15	89.982004	2018-05-05	5	2	5
...
142519	1569	495	1320000	22	0	2	3.86	60000.0	22	569.948187	2018-11-04	11	4	4
142520	1569	474	1200000	15	0	1	65.22	80000.0	15	22.999080	2018-11-04	11	4	4
142521	1569	524	1060000	8	0	3	9.20	132500.0	8	86.956522	2018-11-04	11	4	4
142522	1569	529	600000	5	0	2	5.00	120000.0	5	100.000000	2018-11-04	11	4	4
142523	1569	486	250000	5	0	1	1.79	50000.0	5	279.329609	2018-11-04	11	4	4

142524 rows x 14 columns

Gambar 2. Hasil Analisis Data pada Dataset Penjualan Tiket Film

```
[ ] #Mengetahui info dari deskripsi data
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142524 entries, 0 to 142523
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   film_code              142524 non-null int64
1   cinema_code           142524 non-null int64
2   total_sales            142524 non-null int64
3   tickets_sold           142524 non-null int64
4   tickets_out            142524 non-null int64
5   show_time              142524 non-null int64
6   occu_perc              142399 non-null float64
7   ticket_price           142524 non-null float64
8   ticket_use             142524 non-null int64
9   capacity               142399 non-null float64
10  date                   142524 non-null object
11  month                  142524 non-null int64
12  quarter                142524 non-null int64
13  day                    142524 non-null int64
dtypes: float64(3), int64(10), object(1)
memory usage: 15.2+ MB
```

Gambar 3. Deskripsi Dataset Awal

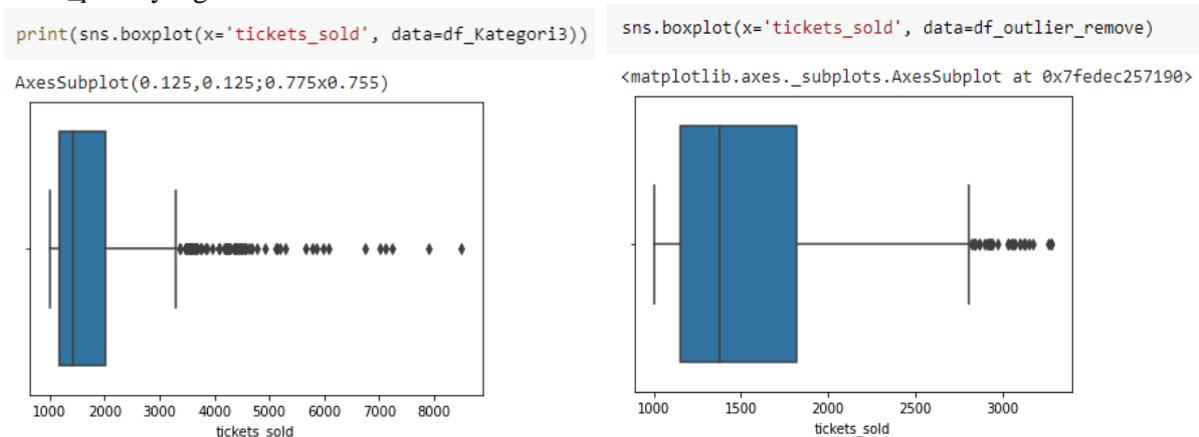
4.2. Hasil Proses Data Preparation

Pada tahapan Data Preparation untuk dataset penjualan tiket film *cinemaTicket-Ref.csv* dilakukan proses pembersihan data agar pada saat dilakukan proses modelling data agar lebih optimal untuk pemodelan data. Tahapan yang dilakukan meliputi pengolahan missing value, proses mengurangi outlier, dan korelasi antar data. Pada pengolahan missing value, dengan menggunakan fungsi *dropna*, maka data yang memiliki nilai kosong akan dihapus dan jumlah data akan direduksi dengan hasil data yang lebih baik. Berdasarkan proses missing value pada gambar 4, dari 142.524 row data terdapat beberapa data yang kosong, dan perlu dihapus. Sehingga setelah menggunakan fungsi *dropna* untuk menghapus nilai kosong, jumlah data yang dihasilkan adalah 142.399 row data.

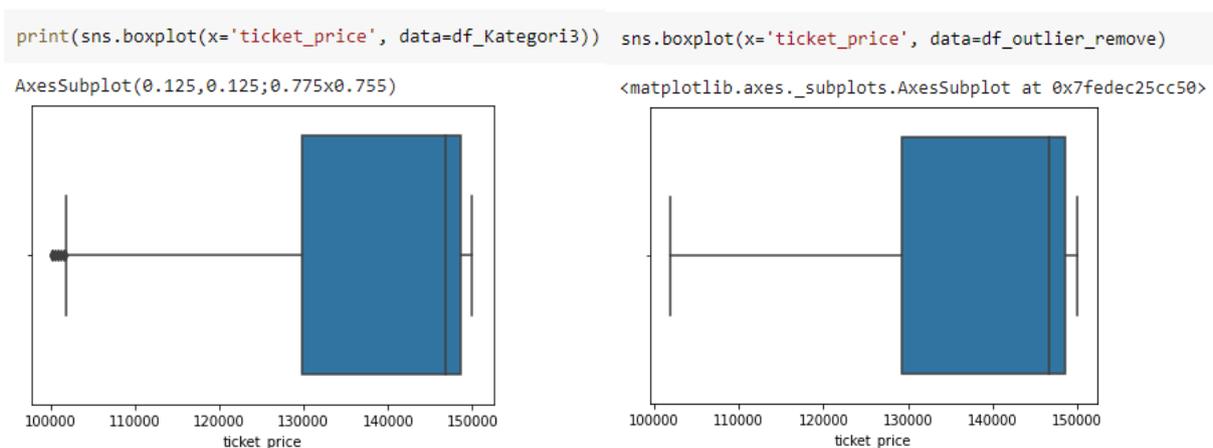


Gambar 4. Proses Penghapusan Missing Value dengan Dropna

Setelah dilakukan proses pengolahan missing value, kemudian proses pengecekan nilai outlier dilakukan guna untuk mencegah terjadinya distorsi data yang dikarenakan adanya nilai yang jauh dari titik pusat data. Pada dataset Penjualan tiket film ini, pada atribut *ticket_sold* dan *ticket_price* terdapat beberapa data yang perlu dilakukan proses penghapusan data outlier, sehingga perlu dilakukan pembersihan data. Berdasarkan gambar 5, pada atribut *ticket_sold* memiliki nilai outlier lebih dari 8000 yang harus dilakukan proses penghapusan data outlier. Begitu pula pada gambar 6 untuk atribut *ticket_price* yang memiliki nilai outlier lebih dari 100000.



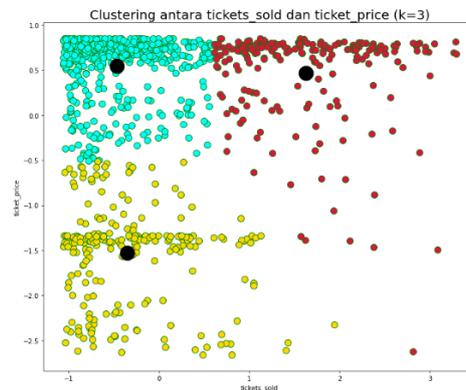
Gambar 5. Proses Penghapusan Outlier untuk Atribut *ticket_sold*



Gambar 6. Proses Penghapusan Outlier untuk Atribut *ticket_price*

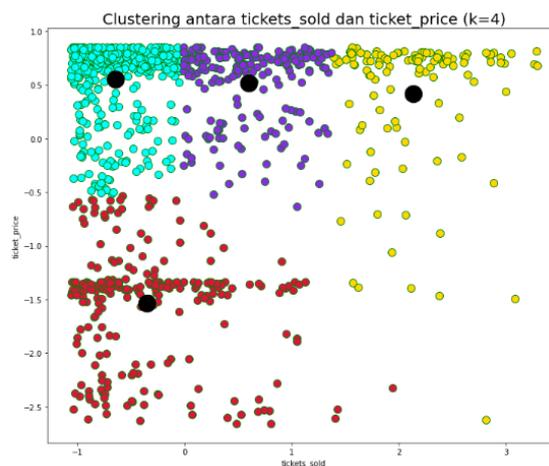
4.3. Hasil Modelling Pengelompokan Tiket dengan K-Means

Pada penelitian ini, dilakukan proses pengelompokan penjualan tiket berdasarkan penjualan dan harga tiket yang ditetapkan dengan menggunakan metode K-Means Clustering. Percobaan pengelompokan untuk penentuan kelompok menggunakan 3 sampai dengan 5 jenis pengelompokan dengan menentukan nilai $k=3$ sampai dengan $k=5$. Berikut adalah hasil dari pengelompokan dari K-Means modelling untuk pengelompokan penjualan tiket dengan nilai $k=3$ sampai dengan nilai $k=5$ pada gambar 7 sampai dengan gambar 9.



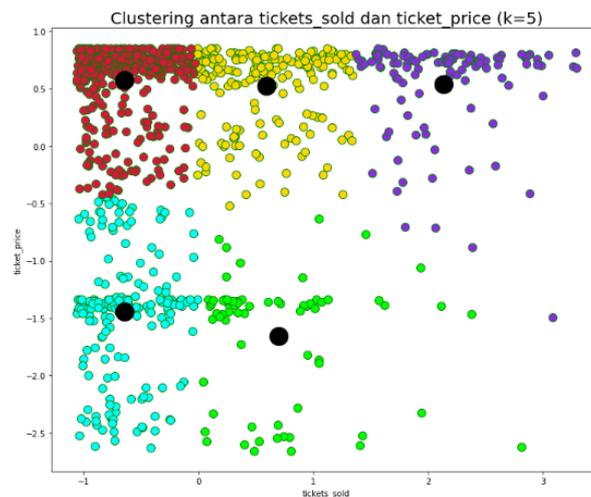
Gambar 7. Hasil K-Means Cluster untuk pengelompokan 3 kategori penjualan tiket

Pada hasil K-Means untuk pengelompokan 3 kategori dari gambar 7, diperoleh hasil bahwa pengelompokan penjualan tiket film berwarna merah dikategorikan sebagai kelompok untuk *ticket_sold* yang terjual dengan jumlah sedang sampai banyak dengan *ticket_price* yang mahal. Selain itu juga, nilai pengelompokan yang berwarna merah hampir sama dengan pengelompokan penjualan tiket yang berwarna biru, yang dimana pengelompokan berwarna biru menunjukkan bahwa *tickets_sold* yang terjual dengan jumlah sedikit sampai sedang dan memiliki *ticket_price* yang mahal.



Gambar 8. Hasil K-Means Cluster untuk pengelompokan 4 kategori penjualan tiket

Pada hasil K-Means untuk pengelompokan 4 kategori pada gambar 8, terdapat 4 pengelompokan penjualan tiket yang ditandai oleh 4 area warna. Warna merah menunjukkan area penjualan tiket film yang paling sedikit dengan harga tiket yang paling rendah. Sedangkan warna biru muda, ungu, dan emas menunjukkan area penjualan tiket dari hasil penjualan paling sedikit ke besar terhadap harga tiket yang mahal. Untuk hasil clustering 4 kelompok ini, hasil cluster berwarna biru muda, ungu, dan emas memiliki nilai yang hampir sama untuk pengelompokan *ticket_price*.

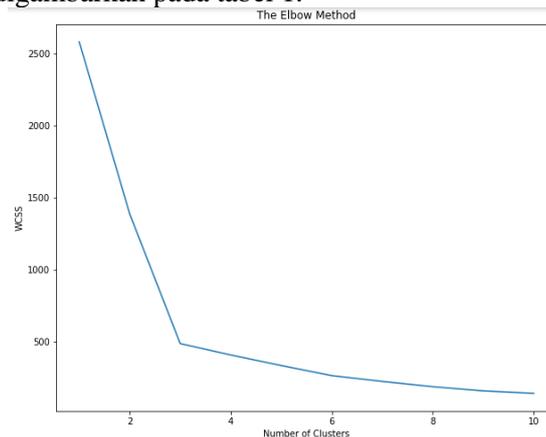


Gambar 9. Hasil K-Means Cluster untuk pengelompokan 5 kategori penjualan tiket

Pada hasil K-Means untuk pengelompokan 5 kategori pada gambar 9 untuk penjualan tiket yang ditandai oleh 5 area warna. Berdasarkan dari pembagian pengelompokan penjualan tiket, jika dilihat dari atribut *ticket_price*, pengelompokan data berwarna emas, ungu, dan merah memiliki nilai dan sebaran data yang hampir sama. Sedangkan jika dilihat dari atribut *ticket_sold*, pengelompokan data berwarna merah dan biru memiliki nilai dan sebaran data yang hampir sama.

4.4. Hasil Evaluasi

Pengukuran evaluasi pengelompokan penjualan tiket film ini dilakukan menggunakan Metode Elbow dan Silhouette Score untuk melihat nilai k untuk pembagian cluster yang paling baik. Berikut adalah hasil dari Metode Elbow untuk penentuan pengelompokan yang ditampilkan pada gambar 10 dan hasil dari Silhouette Score untuk melihat dari penentuan pengelompokan nilai k, manakah yang menjadi nilai k terbaik yang digambarkan pada tabel 1.



Gambar 10. Hasil Penentuan Nilai k dengan *Elbow Method*

Tabel 1. Hasil *Silhouette Score*

No	Nilai k	Nilai Silhouette Score
1.	3	0,42
2.	4	0,51
3.	5	0,60

Berdasarkan hasil *Metode Elbow* untuk penentuan nilai k pada gambar 10, diperoleh hasil bahwa nilai k yang baik untuk penentuan clustering pada data penjualan tiket film ini berada pada nilai $k=3$ sampai dengan nilai $k=5$. Sedangkan dengan nilai $k=6$ dan diatas 6, hasil pengelompokkan akan menjadi tidak seimbang dan jarak hasil cluster akan menjadi lebih jauh dari titik pusat data. Sedangkan pada hasil *Silhouette Score* untuk kasus pengelompokkan penjualan tiket film, diperoleh hasil bahwa nilai k untuk penentuan pengelompokkan dari 3 kelompok sampai dengan 5 kelompok, nilai $k=5$ memiliki hasil pengelompokkan yang lebih baik dibandingkan dengan 2 pengelompokkan lainnya. Untuk nilai dari *Silhouette Score* ini, semakin besar nilai dari *Silhouette*, maka hasil dari pengelompokkan akan lebih baik.

5 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, diambil kesimpulan bahwa penentuan nilai k dapat mempengaruhi pembagian pengelompokkan data pada algoritma K-Means Clustering. Pada data penjualan tiket film, telah dilakukan pengeleompokkan berdasarkan nilai $k=3$ sampai dengan nilai $k=5$. Diperoleh hasil pengelompokkan yang dimana hasil pengelompokkan tidak jauh berbeda jika dilakukan proses pengelompokkan untuk atribut *ticket_sold* dan *ticket_price*. Setelah dilakukan proses evaluasi dengan Elbow Method dan Silhoutte Score, diperoleh hasil bahwa nilai k yang baik untuk penentuan clustering pada data penjualan tiket film ini berada pada nilai $k=3$ sampai dengan nilai $k=5$. Sedangkan dengan nilai $k=6$ dan diatas 6, hasil pengelompokkan akan menjadi tidak seimbang dan jarak hasil cluster akan menjadi lebih jauh dari titik pusat data. Sedangkan pada hasil *Silhouette Score* untuk kasus pengelompokkan penjualan tiket film, diperoleh hasil bahwa nilai k untuk penentuan pengelompokkan dari 3 kelompok sampai dengan 5 kelompok, nilai $k=5$ memiliki hasil pengelompokkan yang lebih baik dibandingkan dengan 2 pengelompokkan lainnya.

Referensi

- [1] N. Astuti, J. N. Utamajaya, and A. Pratama, "Penerapan Data Mining Pada Penjualan Produk Digital Konter Leppangeng Cell Menggunakan Metode K-Means Clustering," *J. Ris. Komput.*, vol. 9, no. 3, pp. 754–760, 2022, doi: 10.30865/jurikom.v9i3.4351.
- [2] Y. D. Yulianti, R. Nurcahyo, and M. Dachyar, "Strategi Penentuan Harga dan Penjualan Hotel Pada Online Travel Agency (OTA)," *Semin. dan Konf. Nas. IDEC*, no. 1, pp. 1–11, 2019.
- [3] A. Batubara and R. Hidayat, "Pengaruh Penetapan Harga dan Promosi terhadap Tingkat Penjualan Tiket pada PSA Mihin Lanka Airlines," *Ilman*, vol. 4, no. 1, pp. 33–36, 2016.
- [4] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, pp. 1–6, 2018, doi: 10.1088/1757-899X/336/1/012017.
- [5] A. Zakir, M. S. Bayudi, and N. Nurjamiah, "Sistem Pendukung Keputusan Evaluasi Loker Pemesanan Tiket Po Medan Jaya Menggunakan Metode K-Means," *J. Teknol. ...*, vol. 3, no. 48, pp. 485–488, 2020, [Online]. Available: <http://jurnal.unprimdn.ac.id/index.php/JUTIKOMP/article/view/1376>.
- [6] R. Pormes and D. H. F. Manongga, "Pemanfaatan Metode Clustering untuk melihat pola penjualan dan perilaku pembelian konsumen, pada penjualan tiket pesawat PT. Garuda Indonesia, Cabang Batam," *J. Tek. Inform. dan Sist. Inf.*, vol. 3, no. 3, pp. 483–493, 2017, doi: 10.28932/jutisi.v3i3.663.
- [7] F. D. S. Alhamdani, A. A. Dianti, and Y. Azhar, "Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode K-Means Clustering," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 2, pp. 70–77, 2021, doi: 10.14421/jiska.2021.6.2.70-77.
- [8] N. H. Harani, C. Prianto, and F. A. Nugraha, "Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python," *J. Manaj. Inform.*, vol. 10, no. 2, pp. 133–146, 2020, doi: 10.34010/jamika.v10i2.2683.
- [9] M. Billah, M. A. Zartesyia, and D. S. Prasvita, "Penerapan Collaborative Filtering, PCA, dan K-Means dalam Pembangunan Sistem Rekomendasi Film," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 579–587, 2021.

- [10] M. Billah, M. A. Zartesyia, and D. S. Prasvita, "Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Ongoing dan Upcoming Film Animasi Jepang," *Senamika*, no. April, pp. 579–587, 2021, [Online]. Available: <https://conference.upnvj.ac.id/index.php/senamika/article/view/1343><https://conference.upnvj.ac.id/index.php/senamika/article/download/1343/1091>.
- [11] N. N. F. R, D. S. Anggraeni, and U. Enri, "Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma K-Means dengan Silhouette Coefficient," *J. Teknol. Inf. Komun.*, vol. 5, no. 2, pp. 29–35, 2022.
- [12] J. Ha, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3th ed. United States of America: Elsevier, 2011.
- [13] R. Llet, M. C. Ortiz, L. A. Sarabia, and M. S. Sánchez, "Selecting variables for k -means cluster analysis by using a genetic algorithm that optimises the silhouettes," *Elsevier*, vol. 515, pp. 87–100, 2004, doi: 10.1016/j.aca.2003.12.020.
- [14] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step Data Mining Guide," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72><http://www.crisp-dm.org/CRISPWP-0800.pdf>.
- [15] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J Multidiscip. Sci. J.*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.