ISSN (Online) 2747-0563
Volume 3 Tahun 2023

Clustering RFM (Recency, Frequency, Monetary) Publisher Gim Menggunakan Algoritma K-Means

Humam Maulana Tsubasanofa Ramadhan[1], Aisyah Pertiwi[2], Galan Ahmad Defanka[3], Anggraini Puspita Sari[4*]

[1,2,3,4] Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
[1]20081010084@student.upnjatim.ac.id
[2]20081010083@student.upnjatim.ac.id
[3]20081010031@student.upnjatim.ac.id
Corresponding author email: anggraini.puspita.if@upnjatim.ac.id

***Abstrak*— *Clustering* merupakan salah satu metode *unsupervised machine learning* yang sangat bermanfaat dalam berbagai bidang, termasuk industri permainan gim yang mengalami pertumbuhan pesat. Dalam industri gim, para penerbit gim memerlukan pemahaman yang lebih baik mengenai pola penjualan gim untuk mengembangkan strategi pemasaran yang efektif. Dalam artikel ini, dilakukan *clustering* data penjualan gim berdasarkan model RFM (*Recency*, *Frequency*, *Monetary*) yang dimiliki penerbit dengan menggunakan algoritma K-Means. Algoritma K-Means digunakan untuk menganalisis data penjualan gim dan mengidentifikasi pola perilisan gim yang sukses, serta karakteristik penerbit yang berkontribusi pada pendapatan yang tinggi. Dengan menerapkan *clustering* pada data penjualan, kita dapat mengidentifikasi kelompok penerbit berdasarkan keaktifan serta banyaknya pendapatan yang didapat oleh penerbit. Dengan demikian, penerbit gim dapat meningkatkan retensi pelanggan, meningkatkan pendapatan, dan mengoptimalkan investasi dalam pengembangan gim baru. Dengan menggunakan pendekatan RFM dan algoritma K-Means, penelitian ini berpotensi memberikan pandangan yang lebih mendalam tentang tren penjualan gim dan faktor-faktor yang mempengaruhi keberhasilan suatu gim di pasar. Dari penelitian ini didapatkan bahwa jumlah atau frekuensi perilisan gim memiliki pengaruh besar terhadap penjualan gim.**

Kata Kunci— **machine learning, clustering, K-Means, gim.**

I. PENDAHULUAN

Industri gim merupakan salah satu sektor yang mengalami pertumbuhan pesat dalam beberapa tahun terakhir. Perkembangan teknologi dan popularitas permainan video telah menciptakan peluang besar bagi pengembang dan penerbit gim. Dalam industri yang semakin kompetitif, penerbit gim harus memahami perilaku pasar dan pola-pola yang terkait dengan penjualan game.

Pada era digital, para penerbit gim memiliki akses terhadap jumlah data yang besar dan kompleks terkait dengan penjualan dan perilisan game. Data ini mencakup informasi seperti tanggal perilisan, frekuensi perilisan, dan pendapatan yang dihasilkan oleh publisher. Untuk menganalisis data ini secara efektif dan mendapatkan wawasan yang berharga, diperlukan pengelompokan atau klasterisasi data. Pengelompokan atau klasterisasi data penjualan memiliki potensi besar untuk memberikan wawasan berharga kepada penerbit gim. Melalui analisis data penjualan, penerbit game dapat mengidentifikasi pola perilisan game yang sukses, menentukan frekuensi optimal dalam merilis game baru, dan mengenali karakteristik pelanggan yang berkontribusi pada pendapatan yang tinggi. Dengan pemahaman yang lebih mendalam tentang pola-pola ini, penerbit game dapat mengarahkan strategi pemasaran mereka dengan lebih tepat sasaran dan meningkatkan keberhasilan bisnis mereka. Metode K-Means merupakan salah satu metode pengelompokan yang umum digunakan. Metode ini berfungsi untuk mengelompokkan objek-objek berdasarkan kesamaan karakteristik dengan memperhatikan jumlah klaster yang ditentukan sebelumnya [1]. Namun, metode K-Means tidak efektif digunakan pada dataset yang memiliki banyak atribut [2]. Untuk mengatasi masalah tersebut, model RFM (*Recency*, *Frequency*, *Monetary*) digunakan untuk mengurangi jumlah atribut pada dataset menjadi tiga atribut yang mencerminkan karakteristik penerbit gim. Penelitian sebelumnya telah menunjukkan keberhasilan penggunaan metode clustering dalam analisis data penjualan dan segmentasi pelanggan. Dalam penelitian yang berjudul "Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global)" oleh A. T. Widiyanto dan A. Witanti, berhasil mengimplementasikan algoritma K-Means dengan baik untuk mengklasifikasikan pelanggan menjadi empat segmen berdasarkan karakteristik masing-masing pelanggan [3].

Pada artikel kali ini, penulis akan melakukan pengelompokan penerbit gim (*publisher*) berdasarkan RFM (*Recency*, *Frequency*, *Monetary*) dengan menggunakan algoritma K-Means.

II. DASAR TEORI

A. *Clustering*

Clustering, yang juga dikenal sebagai pengelompokan merupakan salah satu metode *unsupervised learning* yang dapat melakukan pengelompokan data menjadi beberapa kelompok data (*cluster*) berdasarkan kesamaan karakteristik (*similarity*) yang dimiliki data [4][5].

B. *K-Means*

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

K-Means adalah salah satu metode *clustering* yang berusaha mengelompokkan data menjadi beberapa kelompok data (*cluster*) sehingga data dengan kesamaan (*similarity*) tinggi akan berada dalam satu *cluster* dan data yang berbeda akan berada pada *cluster* lain [6]. Adapun langkah-langkah dalam proses *clustering* dengan metode K-Means adalah sebagai berikut [7]:

1. Menentukan jumlah *cluster* (k) yang diinginkan
2. Menentukan nilai pusat (*centroid*) secara acak
3. Menghitung jarak antara data dengan *centroid* dengan rumus Euclidean yang bisa dilihat pada persamaan (1):

Distance: $$d(x_i, \mu_j) = \sqrt{\Sigma(x_i, \mu_j)^2} \quad (1)$$

Di mana:
- d = titik data
- x_i = data kriteria
- □$_j$ = *centroid* pada *cluster* ke-j

4. Mengelompokkan data berdasarkan kedekatan data dengan *centroid* dan perbarui nilai *centroid* dengan lokasi pusat *cluster* baru dengan menggunakan persamaan (2):

$$\mu_j(t[+1) = \frac{1}{N_{sj}} \Sigma_{j \in sj} x_j \quad (2)$$

Di mana:
- □$_j$ (t+1) = *centroid* baru pada iterasi ke-(t+1)
- N_{sj} = jumlah data pada cluster sj

5. Mengulangi langkah 2 sampai 4 hingga tidak terjadi perubahan anggota *cluster*

C. Preprocess

Preprocess (praproses) merupakan suatu tahapan di mana data yang akan diproses dibersihkan terlebih dahulu dari data-data yang tidak diinginkan (seperti *misisng value*, *duplicate*, *outliers* dll.) dan diolah (seperti *standardscaler*, *dimensionality reduction*, dll.) supaya dapat menghasilkan hasil *clustering* yang lebih bagus [8][9]. Adapun beberapa *preprocess* yang penulis gunakan adalah sebagai berikut.

1) Missing Value

Missing value adalah nilai-nilai yang hilang dari suatu data yang mungkin terjadi karena proses pengambilan data yang kurang sempurna [10]. Untuk mengatasi *missing value* kita bisa menghapus data yang memiliki *missing value*, melakukan estimasi parameter seperti menggunakan algoritma *Expectation-Maximization*, atau imputasi yaitu menghitung nilai pengganti data yang hilang tersebut [10][11].

2) Duplicate Value

Duplicate value merujuk pada duplikasi data atau adanya data yang identik dalam sebuah dataset. Sama seperti *missing value*, *duplicate value* juga penting untuk diatasi untuk menghasilkan hasil *clustering* yang lebih bagus.

3) Outliers

Outliers dapat diartikan sebagai data yang menyimpang jauh dari data lainnya yang dapat berpengaruh buruk pada terhadap pengambilan kesimpulan pada penelitian [12][13]. *Outliers* dapat diidentifikasi, salah satunya dengan menggunakan metode grafik *boxplot* [13][14]. Untuk mengatasi *outliers* biasa digunakan *metric* IQR (*interquartile range*) dengan langkah sebagai berikut [15].

1. Menghitung nilai rata-rata data
2. Menentukan nilai kuartil 3 (Q3) dan nilai kuartil 1 (Q1)
3. Menghitung nilai IQR (*interquartile range*) sesuai persamaan (3) berikut.

$$\text{IQR} = (Q3 - Q1) \quad (3)$$

Di mana:
- IQR = nilai *interquartile range*
- Q3 = nilai kuartil 3
- Q1 = nilai kuartil 1

4. Menghapus data yang bernilai sesuai syarat

4) Standardscaler

Standard Scaler adalah metode *preprocessing* yang umum digunakan dalam analisis data dan *machine learning*. Metode ini bertujuan untuk melakukan standarisasi fitur dengan menghapus rata-rata dan menskalakan unit varian [16]. Proses ini dilakukan pada setiap fitur pada sampel data. *Preprocessing* menggunakan *Standard Scaler* penting dalam analisis data dan *machine learning* karena mencegah adanya data yang memiliki nilai terlalu besar dibandingkan dengan nilai yang lain [17]. Keberadaan data dengan skala yang berbeda-beda dapat mengakibatkan proses training tidak sesuai dengan harapan. Sebagai contoh, apabila satu fitur memiliki rentang nilai yang jauh lebih besar daripada fitur-fitur lainnya, fitur tersebut mungkin akan mendominasi proses training dan memberikan pengaruh yang tidak seimbang pada hasil model. Oleh karena itu, dengan menggunakan *Standard Scaler*, skala nilai dari setiap fitur dapat disesuaikan agar memiliki distribusi yang serupa.

Proses standarisasi fitur dalam *Standard Scaler* terdiri dari dua tahap utama. Tahap pertama adalah penghapusan rata-rata (mean). Untuk setiap fitur, rata-rata dihitung dengan menjumlahkan semua nilai fitur dalam dataset dan membagi nilai fitur dengan jumlah sampel. Kemudian, nilai rata-rata ini dikurangkan dari setiap nilai fitur dalam dataset, sehingga rata-rata fitur menjadi nol. Tahap kedua adalah menskalakan unit varian (standar deviasi). Standar deviasi mengukur sejauh mana nilai-nilai fitur tersebar dari rata-rata. Standar deviasi adalah akar kuadrat dari varian, yang merupakan rata-rata dari kuadrat jarak setiap nilai fitur dengan nilai rata-rata [18]. Dalam *Standard Scaler*, setelah mendapatkan nilai standar deviasi, setiap nilai fitur dalam dataset dibagi dengan standar deviasi yang sesuai untuk menghasilkan fitur yang telah terstandarisasi. Dengan demikian, setiap fitur memiliki rata-rata nol dan standar deviasi satu.

Penggunaan *Standard Scaler* memberikan beberapa manfaat dalam analisis data dan *machine learning*. Pertama, *Standard Scaler* membantu dalam mempercepat konvergensi algoritma *machine learning*. Ketika fitur-fitur memiliki skala nilai yang berbeda-beda, algoritma mungkin memerlukan lebih banyak iterasi untuk mencapai konvergensi yang baik. Dengan menggunakan *Standard Scaler*, fitur-fitur tersebut memiliki skala yang serupa, sehingga konvergensi algoritma dapat tercapai lebih cepat.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

Kedua, *Standard Scaler* membantu dalam mencegah adanya fitur yang memiliki pengaruh yang dominan dalam proses *training*. Jika satu fitur memiliki skala nilai yang jauh lebih besar daripada fitur-fitur lainnya, fitur tersebut mungkin memberikan kontribusi yang tidak proporsional terhadap hasil model. Dengan standarisasi fitur menggunakan Standard Scaler, perbandingan antara fitur-fitur menjadi lebih adil, sehingga mencegah adanya informasi yang salah atau tidak akurat yang diberikan kepada model.

5) Dimensionality Reduction t-SNE

Dimensionality reduction adalah teknik yang penting dalam analisis data dan *machine learning* untuk mengurangi dimensi atau jumlah fitur dalam dataset. Salah satu metode *dimensionality reduction* yang populer adalah t-SNE (*t-Distributed Stochastic Neighbor Embedding*). t-SNE adalah metode *non-linear* yang digunakan untuk memvisualisasikan data dengan dimensi tinggi dalam ruang dua atau tiga dimensi. t-SNE mengatasi masalah dalam memahami hubungan dan struktur data kompleks dengan cara memetakan setiap sampel ke ruang rendah-dimensi, di mana sampel-sampel yang memiliki kesamaan dalam ruang asli akan tetap berdekatan dalam ruang rendah-dimensi. t-SNE bekerja dengan cara memodelkan probabilitas distribusi kedekatan antara pasangan-pasangan sampel dalam ruang asli dan ruang rendah-dimensi. Keuntungan utama dari t-SNE adalah kemampuannya dalam mempertahankan struktur dan hubungan yang kompleks dalam data asli ketika ditransformasikan ke ruang rendah-dimensi. Hal ini memungkinkan visualisasi yang lebih baik dan pemahaman yang lebih dalam tentang data yang memiliki dimensi tinggi.

D. RFM (Recency, Frequency, Monetary)

RFM (*Recency*, *Frequency*, *Monetary*) adalah salah satu analisis perilaku pelanggan berdasarkan tiga variabel yaitu *recency* (terakhir melakukan transaksi), *frequency* (frekuensi transaksi), dan *monetary* (jumlah pengeluaran transaksi setiap pelanggan) [19].

E. Silhouette Score

Silhouette score atau *silhouette index* merupakan salah satu metode validitas *clustering* berbasis internal yang didasarkan pada penilaian susunan objek di setiap kelompok dengan membandingkan jarak rata-rata antar objek di dalam kelompok yang sama atau dengan kelompok lain [20]. Adapun persamaan (4) berikut adalah rumus untuk mendapatkan nilai SI [20][21].

$$SI = \frac{1}{N}\sum_{i=1}^{n} \blacksquare \left(\frac{b(i)-a(i)}{max\{a(i),b(i)\}}\right) \quad (4)$$

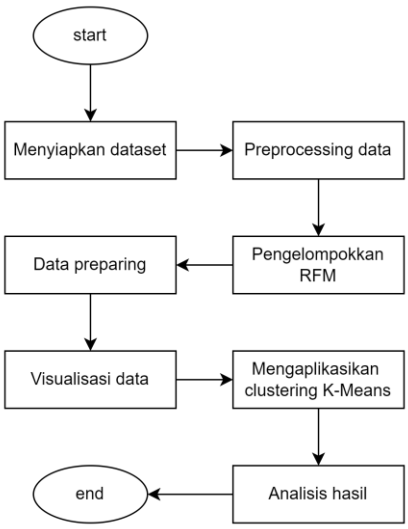
Di mana:

- SI = *silhouette index* atau *silhouette score*
- α(i) = jarak rata-rata sampel i dengan sampel lain di *cluster* yang sama
- b(i) = jarak minimum sampel dengan sampel i ke *cluster* lain

Interpretasi dari *Silhouette Score* adalah sebagai berikut:

- Nilai dekat dengan 1 menunjukkan bahwa pengelompokan sangat baik, dengan sampel-sampel dalam kelompok terpisah secara signifikan dari kelompok lainnya.
- Nilai dekat dengan 0 menunjukkan bahwa pengelompokan sedikit ambigu, dengan beberapa overlap antara kelompok-kelompok.
- Nilai negatif menunjukkan bahwa pengelompokan buruk, dengan banyak sampel yang seharusnya berada dalam kelompok lain.

III. METODOLOGI PENELITIAN



Gambar 1 Alur penelitian

Pada penelitian kali ini, penulis melakukan beberapa tahap yang ditunjukkan dalam Gambar 1, yaitu menyiapkan dataset yang bisa didapat dari situs Kaggle, *preprocessing* data, pengelompokkan RFM, *data preparing*, visualisasi data, mengaplikasikan *clustering* K-Means, dan analisis hasil.

A. *Menyiapkan Dataset*

Dataset yang penulis gunakan adalah data "Video Game Sales" yang penulis dapatkan dari situs Kaggle [22]. Dataset ini berisi 16.598 data judul gim dengan beberapa kolom di dalamnya. Kolom-kolom tersebut antara lain:

1. Rank
   Ranking penjualan gim berdasarkan total penjualan
2. Name
   Nama judul gim
3. Platform
   Platform yang tersedia untuk bisa memainkan gim (seperti PS4, PC, Wii, dan lain-lain.)
4. Year
   Tahun rilis gim
5. Genre
   Genre dari gim (seperti Sports, Racing, Puzzle, dan lain-lain.)
6. Publisher
   Penerbit gim (seperti Nintendo, Sony Computer Entertainment, Electronic Arts, dan lain-lain.)
7. NA_Sales
   Penjualan gim di North America (Amerika Utara) dalam juta dolar

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

8. EU_Sales
   Penjualan gim di Europe (Eropa) dalam juta dolar
9. JP_Sales
   Penjualan gim di Jepang dalam juta dolar
10. Other_Sales
    Penjualan gim di daerah lain dalam juta dolar
11. Global_Sales
    Total penjualan gim di seluruh dunia

B. *Preprocess*

Setelah meng-import dataset ke dalam program Python, dataset terlebih dahulu dilakukan praproses (preprocess) terlebih dahulu guna menghasilkan clustering yang lebih bagus. Praproses yang penulis terapkan pada penelitian ini antara lain:

1. *Missing values*
   Data yang kurang lengkap atau memiliki nilai Null/NaN pada salah satu kolomnya (*missing value*) akan di-drop atau dihapus dari dataset.
2. *Duplicates*
   Data yang identik dengan suatu data lain (*duplicate*) akan dihapus dari dataset dan menyisakan satu data dari data yang identik tadi.
3. Memeriksa anomali lainnya
   Mengecek apakah ada keanehan atau anomali pada dataset. Anomali yang dimaksud adalah seperti rentang nilai minimal dan maksimal, serta tipe data dari setiap kolom pada dataset

C. *Pengelompokan RFM*

Data penjualan gim yang sudah melewati praproses tadi akan dilakukan pengelompokkan RFM (Recency, Frequency, Monetary) berdasarkan masing-masing penerbit (publisher).

1. *Recency*
   Pada artikel ini, *recency* adalah kebaruan atau jarak waktu antara publisher merilis gim terbaru mereka dengan gim terbaru yang dirilis di pasaran.
2. *Frequency*
   Pada artikel ini, *frequency* adalah seberapa sering *publisher* merilis gim dilihat dari jumlah judul yang dirilis masing-masing *publisher*.
3. *Monetary*
   Pada penelitian ini, *monetary* adalah total pendapatan dari semua gim yang dirilis masing-masing *publisher*.
   Dari proses pengelompokan RFM ini dihasilkan dataframe baru yang berisi data *publisher* beserta *recency*, *frequency*, dan *monetary*-nya.

D. *Data Preparing*

Pada tahap ini, data RFM yang didapat dari proses sebelumnya akan diproses lagi sebelum mengaplikasikan algoritma clustering supaya bisa mendapatkan hasil yang maksimal. Proses tersebut antara lain:

1. Mengurangi *outliers*
   Dalam data RFM masih terdapat banyak *outliers* yang dapat mengurangi performa algoritma *clustering*. Pada artikel ini, penulis melakukan pengurangan *outliers* dengan metode IQR.
2. *Standardscaler*
   Standardscaler diaplikasikan untuk melakukan standarisasi fitur-fitur pada dataset. Dari proses standarisasi ini fitur-fitur akan memiliki nilai rata-rata = 0 dan variansi = 1. Standarisasi pada penelitian ini penting karena algoritma K-Means sensitif terhadap skala data.
3. *Dimensionality Reduction*
   Data RFM yang memiliki tiga fitur (*recency*, *frequency*, dan *monetary*) akan dilakukan pengurangan dimensi menjadi dua dimensi dengan metode t-SNE. Proses pengurangan dimensi dilakukan untuk mengurangi kompleksitas, memperbaiki performa algoritma, dan memudahkan penulis ketika ingin melakukan visualisasi data.

E. *Visualisasi Data*

Pada tahap ini dilakukan visualisasi data untuk menampilkan persebaran data sebelum dilakukan *clustering* dengan menggunakan *scatter plot*. Visualisasi data ini nanti bisa digunakan untuk melihat bagaimana *clustering* terjadi dengan melihat grafik persebaran sebelum dan setelah dilakukan *clustering*.

F. *Mengaplikasikan Algoritma K-Means*

Untuk mengaplikasikan algoritma K-Means, penulis perlu menentukan banyak *cluster* yang diinginkan. Penulis mencari banyak *cluster* terbaik dengan menggunakan bantuan dua metode, yaitu *elbow graph* dan *silhouette score*. Setelah ditentukan jumlah *cluster* yang diinginkan, kemudian diaplikasikan *clustering* dengan algoritma K-Means dan membuat dataframe baru dengan tambahan fitur cluster yang didapat dari proses *clustering* dengan algoritma K-Means.

G. *Analisis Hasil*

Pada tahap ini, dilakukan analisis hasil *cluster* dari algoritma K-Means beserta kriteria yang memenuhi pengelompokan *clustering* yang didapat.

IV. HASIL DAN PEMBAHASAN

A. *Dataset*

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...	...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	16597	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows × 11 columns

Gambar 2 Hasil impor dataset

Dataset yang berhasil diimpor menunjukkan terdapat 16.598 data dengan 11 kolom fitur yaitu '*Rank*', '*Name*', '*Platform*', '*Year*', '*Genre*', '*Publisher*', '*NA_Sales*', '*EU_Sales*', '*JP_Sales*', '*Other_Sales*', dan '*Global_Sales*' yang ditunjukkan dalam Gambar 2.

B. *Praproses*

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

Supaya dapat menghasilkan clustering yang lebih baik, dilakukanlah preprocessing data. Berikut adalah hasil dari preprocessing data yang penulis terapkan pada dataset awal.

1) Missing Value

Preprocess awal yang penulis lakukan adalah mengecek apakah ada data yang hilang atau kurang lengkap (*missing values*) pada dataset.

```
# cek missing values
game_df.isnull().sum()
```

```
Rank            0
Name            0
Platform        0
Year          271
Genre           0
Publisher      58
NA_Sales        0
EU_Sales        0
JP_Sales        0
Other_Sales     0
Global_Sales    0
dtype: int64
```

Gambar 3 Hasil cek missing value

Setelah dilakukan pengecekan, terdapat 271 *missing values* di kolom 'Year' dan 58 *missing values* di kolom 'Publisher' ditunjukkan dalam Gambar 3. Missing values tersebut akan dihapus (*drop*) dari dataset.

```
# drop missing values
game_df = game_df.dropna()
game_df.isnull().sum()
```

```
Rank            0
Name            0
Platform        0
Year            0
Genre           0
Publisher       0
NA_Sales        0
EU_Sales        0
JP_Sales        0
Other_Sales     0
Global_Sales    0
dtype: int64
```

Gambar 4 Setelah drop missing value

Setelah dihapus (*drop*) dari dataset, dapat dilihat sudah tidak ada lagi *missing values* pada dataset yang ditunjukkan dalam Gambar 4.

2) Duplicates

Preprocess berikutnya penulis menghapus apabila dalam dataset terdapat data identik (*duplicates*).

```
# shape awal
print(f"shape awal: {game_df.shape}")
# drop duplicates
game_df = game_df.drop_duplicates(keep='first')
# shape setelah
print(f"shape setelah: {game_df.shape}")
```

```
shape awal: (16291, 11)
shape setelah: (16291, 11)
```

Gambar 5 Drop duplicate

Pada Gambar 5 dapat terlihat bahwa *shape* dataset tidak mengalami perubahan, yang artinya tidak terdapat adanya data duplikat pada dataset tersebut.

3) Anomali Lainnya

Pada tahap ini penulis melakukan dua pengecekan yaitu pengecekan rentang nilai dan pengecekan tipe data.

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Gambar 6 Cek rentang nilai

Dari Gambar 6 terlihat semua data memiliki rentang nilai yang normal. Penjualan minimal di masing-masing daerah ('NA_Sales', 'EU_Sales', 'JP_Sales', dan 'Other_Sales') mungkin untuk bernilai 0 karena artinya tidak ada penjualan suatu gim di daerah tersebut. Yang perlu diperhatikan adalah 'Global_Sales' tidak mungkin memiliki nilai minimal 0. Dan seperti yang terlihat pada Gambar 6, 'Global_Sales' memiliki nilai minimal 0,01.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16291 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16291 non-null  int64
 1   Name          16291 non-null  object
 2   Platform      16291 non-null  object
 3   Year          16291 non-null  float64
 4   Genre         16291 non-null  object
 5   Publisher     16291 non-null  object
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.5+ MB
```

Gambar 7 Cek tipe data

Dari Gambar 7 terlihat data 'Year' memiliki tipe data *float64*. Seharusnya data 'Year' memiliki tipe data *int64*. Oleh karena

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

itu penulis melakukan pengubahan tipe data 'Year' dari *float64* menjadi *int64*.

```
# mengganti Year data type
game_df['Year'] = pd.to_datetime(game_df['Year'], format='%Y').dt.year
game_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16291 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16291 non-null  int64
 1   Name          16291 non-null  object
 2   Platform      16291 non-null  object
 3   Year          16291 non-null  int64
 4   Genre         16291 non-null  object
 5   Publisher     16291 non-null  object
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(5), int64(2), object(4)
memory usage: 1.5+ MB
```

Gambar 8 Setelah ganti tipe data

Dari Gambar 8 bisa terlihat bahwa tipe data 'Year' sudah berhasil diubah dari *float64* menjadi *int64*.

C. *Pengelompokan RFM (Recency, Frequency, Monetary)*

1) Recency

Untuk mencari *recency*, penulis perlu mencari tahun rilis terbaru yang ada pada dataset.

```
# menacri max_date (tahun rilis game terbaru di dataset)
max_date = max(game_df['Year'])
max_date
```

```
2020
```

Gambar 9 Tahun rilis terbaru

Dari Gambar 9 didapat bahwa tahun rilis gim terbaru pada dataset adalah 2020. Kemudian penulis menghitung jarak tahun rilis tiap game dengan tahun 2020 (tahun rilis gim terbaru) dan menyimpannya pada kolom 'Year_Diff'.

```
# menghitung perbedaan tahun rilis setiap game
game_df['Year_Diff'] = max_date - game_df['Year']
game_df
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Year_Diff
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74	14
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	35
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82	12
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00	11
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	24
...	...	...	...	...	...	...	...	...	...	...	...	...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002	Platform	Kemco	0.01	0.00	0.00	0.00	0.01	18
16594	16597	Men in Black II: Alien Escape	GC	2003	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01	17
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008	Racing	Activision	0.00	0.00	0.00	0.00	0.01	12
16596	16599	Know How 2	DS	2010	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01	10
16597	16600	Spirits & Spells	GBA	2003	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01	17

16291 rows × 12 columns

Gambar 10 Mencari jarak tahun rilis setiap gim

Setelah didapatkan jarak tahun rilis setiap gim. Selanjutnya penulis membuat dataframe baru yang berisi data *publisher* dan jarak tahun rilis gim terbaru dari masing-masing *publisher*.

	Publisher	Recency
0	10TACLE Studios	13
1	1C Company	9
2	20th Century Fox Video Games	38
3	2D Boy	12
4	3DO	17
...	...	...
571	id Software	28
572	imageepoch Inc.	6
573	inXile Entertainment	5
574	mixi, Inc	5
575	responDESIGN	15

576 rows × 2 columns

Gambar 11 Dataframe recency publisher

Gambar 11 adalah dataframe *recency publisher*. Didapatkan 576 data *publisher* yang terdapat pada dataset penjualan gim.

2) Frequency

Untuk mendapatkan *frequency*, penulis menghitung berapa banyak jumlah judul (kolom 'Name') yang dirilis tiap *publisher* dan disimpan pada dataframe *frequency publisher*.

	Publisher	Frequency
0	10TACLE Studios	3
1	1C Company	3
2	20th Century Fox Video Games	5
3	2D Boy	1
4	3DO	36
...	...	...
571	id Software	1
572	imageepoch Inc.	2
573	inXile Entertainment	1
574	mixi, Inc	1
575	responDESIGN	2

576 rows × 2 columns

Gambar 12 Dataframe frequency publisher

Gambar 12 adalah dataframe *frequency* publisher. Didapatkan 576 data publisher dan jumlah (*frequency*) gim yang dirilis.

3) Monetary

Untuk mendapatkan *monetary*, penulis menjumlah total penjualan (kolom 'Global_Sales') tiap *publisher* dan disimpan pada dataframe *monetary publisher*.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

	Publisher	Monetary
0	10TACLE Studios	0.11
1	1C Company	0.10
2	20th Century Fox Video Games	1.94
3	2D Boy	0.04
4	3DO	10.12
...	...	...
571	id Software	0.03
572	imageepoch Inc.	0.04
573	inXile Entertainment	0.10
574	mixi, Inc	0.86
575	responDESIGN	0.13

576 rows × 2 columns

Gambar 13 Dataframe monetary publisher

Gambar 13 adalah dataframe *monetary publisher*. Didapatkan 576 data *publisher* dan total penjualan (*monetary*) yang didapat *publisher* dari gim-gim yang mereka rilis.

4) Menggabungkan Dataframe

Setelah membuat tiga dataframe RFM (*recency*, *frequency*, *monetary*), kemudian penulis menggabungkan ketiga dataframe tersebut menjadi satu dataframe.

	Publisher	Recency	Frequency	Monetary
0	10TACLE Studios	13	3	0.11
1	1C Company	9	3	0.10
2	20th Century Fox Video Games	38	5	1.94
3	2D Boy	12	1	0.04
4	3DO	17	36	10.12
...	...	...	...	...
571	id Software	28	1	0.03
572	imageepoch Inc.	6	2	0.04
573	inXile Entertainment	5	1	0.10
574	mixi, Inc	5	1	0.86
575	responDESIGN	15	2	0.13

576 rows × 4 columns

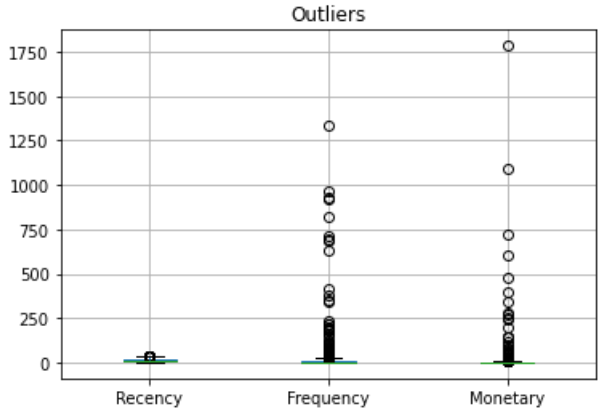
Gambar 14 Dataframe gabungan RFM

Gambar 14 adalah dataframe rfm_df hasil penggabungan RFM (*recency*, *frequency*, *monetary*). Didapatkan 576 data publisher dan nilai RFM dari masing-masing *publisher*.

D. *Data Preparing*

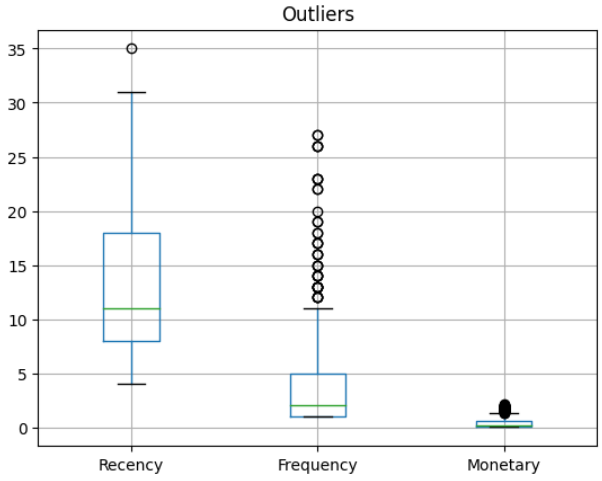
Dataframe rfm_df hasil dari proses sebelumnya masih perlu di-*preprocess* terlebih dahulu sebelum diaplikasikan dengan algoritma clustering.

1) Outliers



Gambar 15 Outlier awal dataframe RFM

Dapat dilihat pada Gambar 15, data RFM yang belum di-*preprocess* masih terdapat sangat banyak outliers di dalamnya. Oleh karena itu, penulis mengurangi jumlah outliers dengan menggunakan metode IQR.



Gambar 16 Outlier setelah dikurangi

Gambar 16 menunjukkan jumlah outliers pada data sudah berkurang signifikan setelah dilakukan pengurangan outliers dengan menggunakan metode IQR.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

	Publisher	Recency	Frequency	Monetary
0	10TACLE Studios	13	3	0.11
1	1C Company	9	3	0.10
3	2D Boy	12	1	0.04
5	49Games	11	1	0.04
8	7G//AMES	9	4	0.08
...	...	...	...	...
571	id Software	28	1	0.03
572	imageepoch Inc.	6	2	0.04
573	inXile Entertainment	5	1	0.10
574	mixi, Inc	5	1	0.86
575	responDESIGN	15	2	0.13

439 rows × 4 columns

Gambar 17 Dataframe setelah outlier dikurangi

Gambar 17 adalah data RFM yang mengalami pengurangan jumlah dari 576 data menjadi 439 data setelah dilakukan pengurangan *outliers*.

2) Standardscaler

Standarisasi dengan *standardscaler* dilakukan untuk mengubah skala variabel sehingga memiliki nilai mean = 0 dan deviasi standar = 1.

	Publisher	Recency	Frequency	Monetary
0	10TACLE Studios	-0.008044	-0.221912	-0.570459
1	1C Company	-0.596608	-0.221912	-0.589913
3	2D Boy	-0.155185	-0.626982	-0.706637
5	49Games	-0.302326	-0.626982	-0.706637
8	7G//AMES	-0.596608	-0.019377	-0.628821
...	...	...	...	...
571	id Software	2.199071	-0.626982	-0.726091
572	imageepoch Inc.	-1.038031	-0.424447	-0.706637
573	inXile Entertainment	-1.185172	-0.626982	-0.589913
574	mixi, Inc	-1.185172	-0.626982	0.888592
575	responDESIGN	0.286238	-0.424447	-0.531551

439 rows × 4 columns

Gambar 18 Dataframe setelah standarisasi

Gambar 18 adalah data RFM yang sudah dilakukan standarisasi dengan *standardscaler*.

3) Dimensionality Reduction

Setelah dilakukan standarisasi, data RFM akan dilakukan *dimensionality reduction* dengan metode t-SNE untuk mengurangi dimensi menjadi dua dimensi.

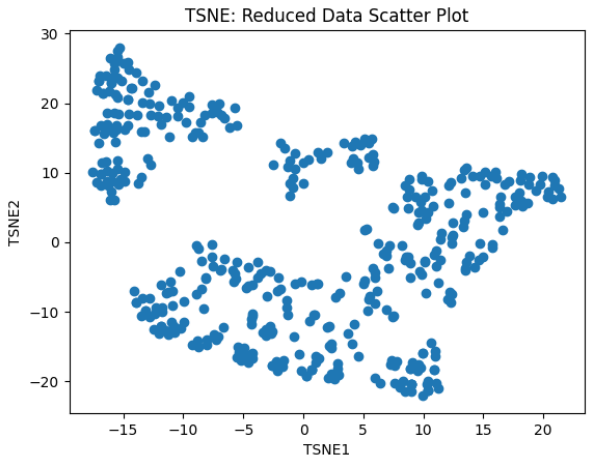
	index	Publisher	Recency	Frequency	Monetary	TSNE1	TSNE2
0	0	10TACLE Studios	13	3	0.11	-5.611387	-5.221938
1	1	1C Company	9	3	0.10	0.525597	-11.419552
2	3	2D Boy	12	1	0.04	-10.996088	-11.464293
3	5	49Games	11	1	0.04	-8.090295	-14.780025
4	8	7G//AMES	9	4	0.08	0.951176	-10.437887
...	...	...	...	...	...	...	...
434	571	id Software	28	1	0.03	-15.485978	26.587389
435	572	imageepoch Inc.	6	2	0.04	7.256232	-17.638596
436	573	inXile Entertainment	5	1	0.10	7.640616	-20.217438
437	574	mixi, Inc	5	1	0.86	7.405691	4.966956
438	575	responDESIGN	15	2	0.13	-10.280675	-4.252523

439 rows × 7 columns

Gambar 19 Dataframe hasil dimensionality reduction

Pada Gambar 19 dapat terlihat kolom ‘TSNE1’ dan ‘TSNE2’ adalah hasil dari proses dimensionality reduction. Pada gambar di atas terlihat ada data asli (masih belum di-*standardscaler*). Data tersebut ada untuk memudahkan penulis dalam memantau perubahan data dan hasil dari tiap prosesnya.

E. *Visualisasi Data*



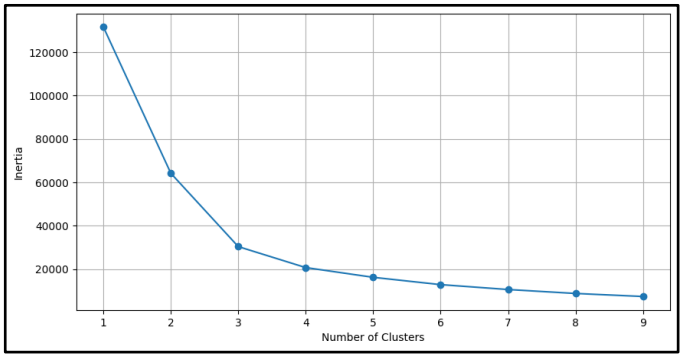
Gambar 20 Scatter plot sebelum clustering

Gambar 20 adalah visualisasi persebaran data (sudah melalui proses dimensionality reduction t-SNE) yang nantinya akan dilakukan *clustering*.

F. *Mengaplikasikan Algoritma K-Means*

Sebelum melakukan *clustering* dengan algoritma K-Means, terlebih dahulu penulis mencari jumlah *cluster* yang terbaik.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023



Gambar 21 Elbow Graph

Gambar 21 adalah *elbow graph*. Dari grafik tersebut sekilas titik siku berada pada angka 3 yang berarti kemungkinan jumlah *cluster* terbaik adalah 3. Namun penulis menggunakan *silhouette score* untuk memastikan lagi.

```
silhouette(reduced_df[['TSNE1','TSNE2']], 10)
```

```
For n_cluster=2, the silhouette score is 0.48193255066871643
For n_cluster=3, the silhouette score is 0.539847731590271
For n_cluster=4, the silhouette score is 0.5189237594604492
For n_cluster=5, the silhouette score is 0.4839951694011688
For n_cluster=6, the silhouette score is 0.49854743480682373
For n_cluster=7, the silhouette score is 0.48472005128860474
For n_cluster=8, the silhouette score is 0.4738781452178955
For n_cluster=9, the silhouette score is 0.48974838852882385
```

Gambar 22 Silhouette Score

Dari Gambar 22 terlihat nilai tertinggi ada pada jumlah cluster 3 dengan nilai 0,5398 yang berarti jumlah *cluster* yang penulis tentukan adalah 3. Setelah menemukan jumlah *cluster*, penulis menerapkan algoritma K-Means untuk dilakukan clustering.

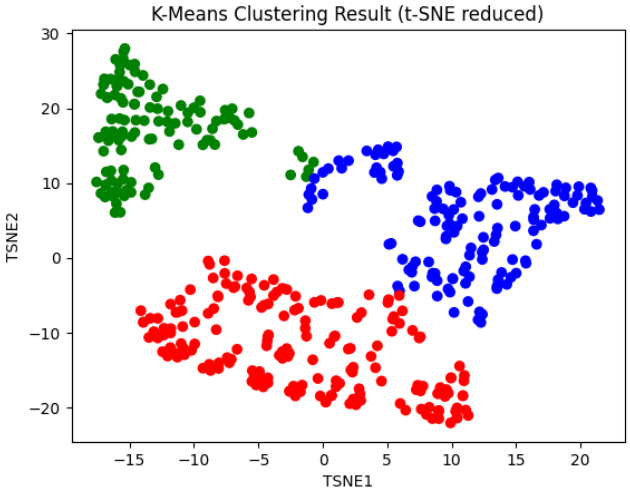
	index	Publisher	Recency	Frequency	Monetary	TSNE1	TSNE2	Cluster
0	0	10TACLE Studios	13	3	0.11	-5.611387	-5.221938	0
1	1	1C Company	9	3	0.10	0.525597	-11.419552	0
2	3	2D Boy	12	1	0.04	-10.996088	-11.464293	0
3	5	49Games	11	1	0.04	-8.090295	-14.780025	0
4	8	7G//AMES	9	4	0.08	0.951176	-10.437887	0
...	...	...	...	...	...	...	...	...
434	571	id Software	28	1	0.03	-15.485978	26.587389	1
435	572	imageepoch Inc.	6	2	0.04	7.256232	-17.638596	0
436	573	inXile Entertainment	5	1	0.10	7.640616	-20.217438	0
437	574	mixi, Inc	5	1	0.86	7.405691	4.966956	2
438	575	responDESIGN	15	2	0.13	-10.280675	-4.252523	0

439 rows × 8 columns

Gambar 23 Dataframe hasil clustering K-Means

Gambar 23 adalah dataframe hasil clustering dengan algoritma K-Means.

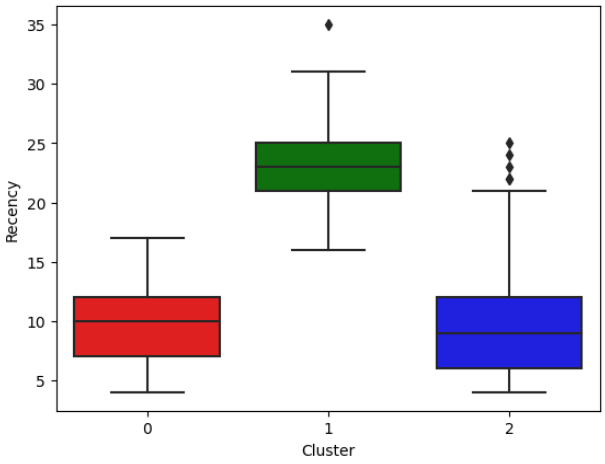
G. *Analisis Hasil*



Gambar 24 Scatter plot setelah clustering

Gambar 24 adalah grafik persebaran data setelah diterapkan K-Means *clustering*. Berikut adalah analisis RFM di setiap cluster hasil K-Means clustering:

1. *Recency*



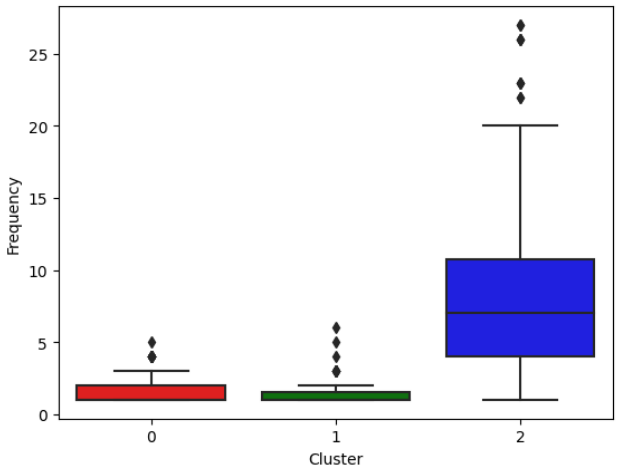
Gambar 25 Boxplot recency

Dari grafik boxplot yang ditunjukkan Gambar 25 didapat bahwa:

- *Publisher* pada cluster0 adalah publisher yang terhitung masih aktif merilis gim dibuktikan dengan dengan recency 4-17 tahun dengan mayoritas berada di rentang 7-12 tahun
- *Publisher* pada cluster1 adalah publisher yang terhitung sudah lama tidak merilis gim dibuktikan dengan dengan recency 16-35 tahun dengan mayoritas berada di rentang 20-25 tahun
- *Publisher* pada cluster2 adalah publisher yang terhitung masih aktif merilis gim dibuktikan dengan dengan recency 4-25 tahun dengan mayoritas berada di rentang 6-12 tahun

2. *Frequency*

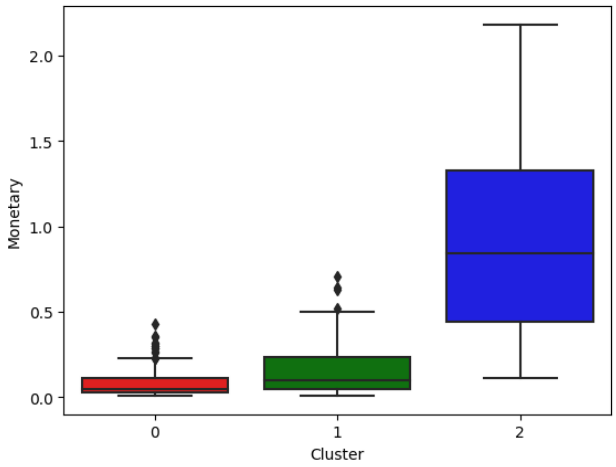
ISSN (Online) 2747-0563
Volume 3 Tahun 2023



Gambar 26 Boxplot frequency

Dari grafik boxplot yang ditunjukkan Gambar 26 didapat bahwa:
- *Publisher* pada cluster0 adalah publisher yang merilis sedikit gim dibuktikan dengan dengan frequency 1-5 judul dengan mayoritas berada di rentang 1-3 judul
- *Publisher* pada cluster1 adalah publisher yang merilis sedikit gim dibuktikan dengan dengan frequency 1-6 judul dengan mayoritas berada di rentang 1-2 judul
- *Publisher* pada cluster2 adalah publisher yang merilis banyak gim dibuktikan dengan dengan frequency 1-27 judul dengan mayoritas berada di rentang 5-12 judul

3. *Monetary*



Gambar 27 Boxplot monetary

Dari Gambar 27 didapat bahwa:
- *Publisher* pada cluster0 adalah publisher yang memiliki pendapatan sangat kecil dibuktikan dengan monetary 0,01-0,48 juta dollar dengan mayoritas berada di rentang 0,03-0,13 juta dollar
- *Publisher* pada cluster1 adalah publisher yang memiliki pendapatan kecil dibuktikan dengan monetary 0,01-0,9 juta dollar dengan mayoritas berada di rentang 0,05-0,27 juta dollar
- *Publisher* pada cluster2 adalah publisher yang memiliki pendapatan besar dibuktikan dengan monetary 0,11-2,18 juta dollar dengan mayoritas berada di rentang 0,49-1,42 juta dollar

V. KESIMPULAN

Dari penelitian yang sudah penulis lakukan dan hasil yang sudah dibahas, dapat ditarik beberapa kesimpulan sebagai berikut:
1. *Clustering* data RFM *publisher* menggunakan algoritma K-Means memiliki nilai *Silhouette Score* atau *Silhouette Index* sebesar 0,539847731590271 yang menunjukkan bahwa hasil pengelompokan yang baik dengan kelompok-kelompok dalam data terpisah secara signifikan dari kelompok lainnya.
2. Data RFM *publisher* berhasil dikelompokkan menjadi tiga kelompok (*cluster*)
3. *Publisher* di cluster0 sebanyak 186 *publisher* memiliki kriteria aktif merilis gim (mayoritas *recency* 7-12 tahun), sedikit merilis gim (mayoritas *frequency* 1-3 judul), dan pendapatan sangat sedikit (mayoritas *monetary* 0,03-0,13 juta dollar)
4. *Publisher* di cluster1 sebanyak 113 *publisher* memiliki kriteria lama tidak aktif merilis gim (mayoritas *recency* 20-25 tahun), sedikit merilis gim (mayoritas *frequency* 1-2 judul), dan pendapatan sedikit (mayoritas *monetary* 0,05-0,27 juta dollar)
5. *Publisher* di cluster2 sebanyak 140 *publisher* memiliki kriteria aktif merilis gim (mayoritas *recency* 6-12 tahun), banyak merilis gim (mayoritas *frequency* 5-12 judul), dan pendapatan banyak (mayoritas *monetary* 0,49-1,42 juta dollar)

UCAPAN TERIMA KASIH

Terima kasih penulis ucapkan kepada Ibu Dr. Eng. Ir, Anggraini Puspita Sari, S.T., M.T. selaku dosen pengampu mata kuliah Machine Learning yang sudah membantu penulis untuk menyelesaikan penelitian ini, serta teman-teman penulis yang sudah berusaha menyelesaikan penelitian ini.

REFERENSI

[1] R. Gustriansyah, N. Suhandi dan F. Antony, "Clustering optimization in RFM analysis based on K-means," Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no. 1, pp. 470-477, 2020.
[2] R. Dash, D. Mishra, A. K. Rath dan M. Acharya, "A hybridized K-means clustering approach for high dimensional dataset," International Journal of Engineering, Science and Technology, vol. 2, no. 2, pp. 59-66, 2010.
[3] A. Widiyanto dan A. Witanti, "Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global)," KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, vol. 1, no. 1, pp. 204-215, 2021.
[4] A. Bastian, H. Sujadi dan G. Febrianto, "Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka)," Jurnal Sistem Informasi, vol. 14, no. 1, pp. 28-34, 2018.
[5] B. M. Metisen dan H. L. Sari, "ANALISIS CLUSTERING MENGGUNAKAN METODE K-MEANS DALAM PENGELOMPOKKAN PENJUALAN PRODUK PADA SWALAYAN FADHILA," Jurnal Media Infotama, vol. 11, no. 2, pp. 110-118, 2015.
[6] S. A. Rahmah, "KLASTERISASI POLA PENJUALAN PESTISIDA MENGGUNAKAN METODE K-MEANS CLUSTERING (STUDI

KASUS DI TOKO JUANDA TANI KECAMATAN HUTABAYU RAJA)," Djtechno : Journal of Information Technology Research, vol. 1, no. 1, pp. 1-5, 2020.
[7] F. Indriyani dan E. Irfiani, "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means," JUITA: Jurnal Informatika, vol. 7, no. 2, pp. 109-114, 2019.
[8] S. F. Pane dan J. Ramdan, "Pemodelan Machine Learning: Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter," Jurnal Sistem Cerdas, vol. 05, no. 01, pp. 12-20, 2022.
[9] M. F. Naufal, Subrata, A. F. Susanto dan C. N. Kansil, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank," Techno.COM, vol. 22, no. 1, pp. 1-11, 2023.
[10] R. Maulid, "Kursus Belajar Data: Mengenal Apa Itu Missing Value," 22 April 2021. [Online]. Available: https://dqlab.id/kursus-belajar-data-mengenal-apa-itu-missing-value. [Diakses 20 Juni 2023].
[11] J. Gifari, "Teknik Pengolahan Data : Mengenal Missing Values dan Cara-Cara Menanganinya," 12 Oktober 2020. [Online]. Available: https://dqlab.id/digital-transformation-pahami-teknik-pengolahan-ini-dalam-industri-data. [Diakses 20 Juni 2023].
[12] D. Hawkins, Identification of Outliers, 1 penyunt., Springer Dordrecht, 2014.
[13] P. R. Sihombing, Suryadiningrat, D. A. Sunarjo dan Y. P. A. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," Jurnal Ekonomi dan Statistik Indonesia, vol. 2, no. 3, pp. 307-316, 2022.
[14] M. F. Triola, Elementary Statistics Using Excel, 7, Penyunt., Pearson, 2021.
[15] M. R. Irianto, A. Maududie dan F. N. Arifin, "Implementation of K-Means Clustering Method for Trend Analysis of Thesis Topics (Case Study: Faculty of Computer Science, University of Jember)," Berkala Sainstek, vol. 10, no. 4, pp. 210-226, 2022.
[16] V. R. Prasetyo, M. Mercifia, A. Averina, L. Sunyoto dan B. Budiarjo, "PREDIKSI RATING FILM PADA WEBSITE IMDB MENGGUNAKAN METODE NEURAL NETWORK," Jurnal Ilmiah NERO, vol. 7, no. 1, pp. 1-8, 2022.
[17] A. Ambarwari, Q. J. Adrian dan Y. Herdiyeni, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritma Machine Learning untuk Identifikasi Tanaman," Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 4, no. 1, pp. 117-122, 2020.
[18] G. Sumantri, M. D. Novianto dan P. P. Prihastuti, "Implementasi Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia untuk Pemerataan Kualitas Pendidikan," Prosiding Seminar Pendidikan Matematika dan Matematika, vol. 8, pp. 1-9, 2023.
[19] W. A. Taqwim, N. Y. Setiawan dan F. A. Bachtiar, "Analisis Segmentasi Pelanggan Dengan RFM Model Pada Pt. Arthamas Citra Mandiri Menggunakan Metode Fuzzy C-Means Clustering," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 2, pp. 1986-1993, 2019.
[20] A. N. Sihananto, A. P. Sari, H. Khariono, R. A. Fernanda dan D. C. M. Wijaya, "Implementasi Metode K-Means Untuk Pengelompokan Kasus Covid-19," Jurnal Informatika dan Sistem Informasi (JIFoSI), vol. 3, no. 1, pp. 76-85, 2022.
[21] R. Adha, N. Nurhaliza, U. Soleha, dan Mustakim, 2021. Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia. Jurnal Sains, Teknologi dan Industri, 18 (2), pp. 206–211.
[22] G. Smith, "Video Game Sales," [Online]. Available: https://www.kaggle.com/datasets/gregorut/videogamesales. [Diakses 4 Juni 2023].
[23]