PREPROCESSING DATASET CSM (CONVENTIONAL AND SOCIAL MEDIA MOVIES) 2014 DAN 2015

Oleh: Hayunitya Edadwi Pratita (23/518670/TK/57134)

DESKRIPSI DATASET

Dataset yang digunakan untuk *preprocessing* adalah dataset CSM (*Conventional and Social Media Movies*) yang diambil dari UCI Machine Learning Repository. Dataset terdiri dari total 231 baris dan 14 kolom untuk periode 2014 dan 2015. Setiap baris mewakili satu judul film pada tahun rilisnya. Data menggabungkan informasi konvensional dari IMDb serta jejak media sosial dari YouTube dan Twitter. Dataset bersifat terbuka, masih ada nilai yang hilang, dan rentang nilainya cukup lebar.

Secara isi, kolom dapat dikelompokkan menjadi dua. Kelompok konvensional yang mencakup Gross sebagai pendapatan, Budget sebagai biaya produksi, Screens sebagai jumlah layar, Ratings sebagai penilaian pengguna, kategori Genre, dan penanda Sequel. Kelompok media sosial memuat Views, Likes, Dislikes, Comments, Aggregate Followers, dan Sentiment. Pada analisis ini juga dibuat dua kolom turunan, yaitu ROI *(Return on Investment)* sebagai efisiensi biaya dan EngagementRate sebagai ukuran interaksi.

RINGKASAN DAN VISUALISASI SETELAH PREPROCESSING

1. Validasi nilai

Tahap ini memastikan nilai berada pada domain yang benar. Tahun dibatasi ke 2014 dan 2015. Ratings harus 0 sampai 10. Semua kolom numerik tidak boleh negatif. Nilai yang melanggar aturan ditandai kosong agar ditangani di tahap missing values. Duplikasi pada pasangan Movie dan Year dihapus bila ada. Hasil dari tahap ini adalah tidak ada baris yang berkurang karena tidak ditemukan duplikasi.

2. Missing values

Tahap ini membersihkan nilai kosong agar perhitungan aman. Nol pada Budget dan Views diperlakukan sebagai kosong. Baris yang kosong pada Gross atau Budget dihapus. Kolom numerik lain yang memiliki baris kosong diisi median per tahun. Hasil dari tahap ini adalah data terhapus 1 baris sehingga data tersisa 230 baris dan 14 kolom. Setelah imputasi, seluruh kolom tidak memiliki nilai kosong.

Missing sebelum:

	0
Movie	0
Year	0
Ratings	0
Genre	0
Gross	0
Budget	1
Screens	10
Sequel	0
Sentiment	36
Views	0
Likes	0
Dislikes	0
Comments	0
Aggregate Followers	35

Missing sesudah:

	0
Movie	0
Year	0
Ratings	0
Genre	0
Gross	0
Budget	0
Screens	0
Sequel	0
Sentiment	0
Views	0
Likes	0
Dislikes	0
Comments	0
Aggregate Followers	0

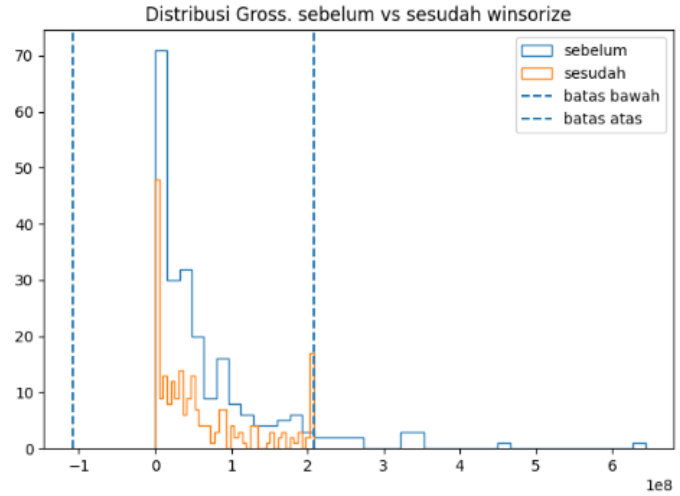
Gambar 1 dan 2. *Missing* sebelum dan sesudah

3. Feature engineering

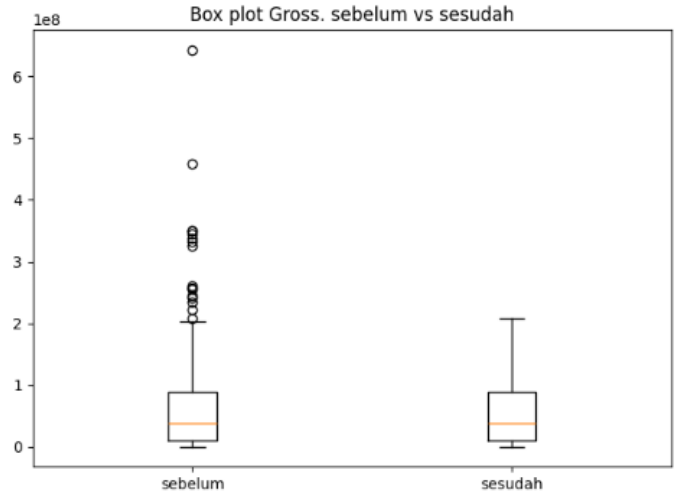
Tahap ini menambah dua kolom turunan agar pembacaan menjadi lebih informatif. ROI dihitung sebagai Gross per Budget, jika tidak maka nilainya dibuat kosong. EngagementRate dihitung sebagai (Likes + Comments) per Views. Hasil dari tahap ini adalah kolom ROI dan EngagementRate ditambahkan. Tidak ada baris yang dihapus. Nilai yang tidak bisa dihitung karena Budget atau Views nol ditandai kosong sehingga aman untuk analisis.

4. Outlier pada Gross dan Budget

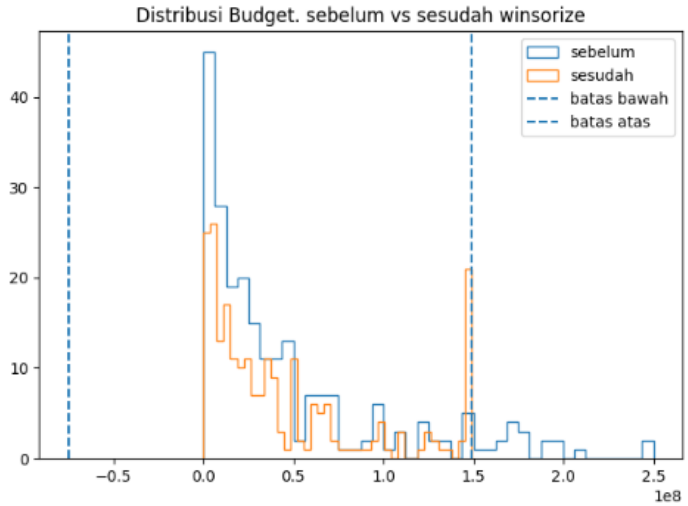
Tahap ini mengendalikan nilai ekstrem agar ringkasan dan grafik tidak bias. Fokus hanya pada kolom Gross dan Budget karena dua kolom ini langsung membentuk ROI. Metode yang digunakan adalah winsorizing berbasis IQR dengan k = 1,5. Batas bawah dihitung Q1 − k × IQR, batas atas Q3 + k × IQR.



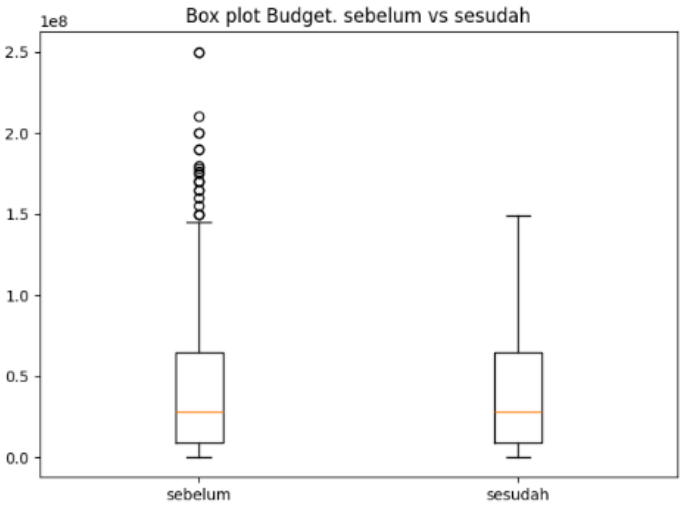
Gambar 3. Histogram Gross sebelum dan sesudah



Gambar 4. Boxplot Gross sebelum dan sesudah



Gambar 5. Histogram Budget sebelum dan sesudah



Gambar 6. Boxplot Budget sebelum dan sesudah

5. Cek redundansi kolom sosial

Tujuan tahap ini adalah menilai apakah ada kolom sosial yang sangat berkorelasi sehingga informasinya tumpang tindih. Metode yang dipakai adalah korelasi Pearson pada Views, Likes, Dislikes, Comments, dan Aggregate Followers.

Korelasi antar metrik sosial:

	Views	Likes	Dislikes	Comments	Aggregate Followers
Views	1.000000	0.676876	0.775796	0.710125	0.159283
Likes	0.676876	1.000000	0.470099	0.917421	0.087796
Dislikes	0.775796	0.470099	1.000000	0.579456	0.061559
Comments	0.710125	0.917421	0.579456	1.000000	0.042352
Aggregate Followers	0.159283	0.087796	0.061559	0.042352	1.000000

Gambar 7. Korelasi antar metrik sosial

```
Pasangan korelasi:
Likes vs Comments: 0.917
Views vs Dislikes: 0.776
Views vs Comments: 0.710
Views vs Likes: 0.677
Dislikes vs Comments: 0.579
Likes vs Dislikes: 0.470
Views vs Aggregate Followers: 0.159
Likes vs Aggregate Followers: 0.088
Dislikes vs Aggregate Followers: 0.062
Comments vs Aggregate Followers: 0.042
```

Gambar 8. Pasangan korelasi

6. Evaluasi kolom Sentiment

Tujuan tahap ini adalah menilai apakah kolom Sentiment akan dipertahankan karena memiliki banyak nilai bernilai nol. Pemeriksaan dilakukan pada tiga hal, yaitu proporsi nilai 0, tingkat variasi, dan korelasi terhadap Gross serta ROI. Kolom akan dihapus jika proporsi 0 mencapai atau melebihi 0,90, atau varians sangat kecil, dan korelasi absolut ke Gross serta ROI kurang dari 0,05.

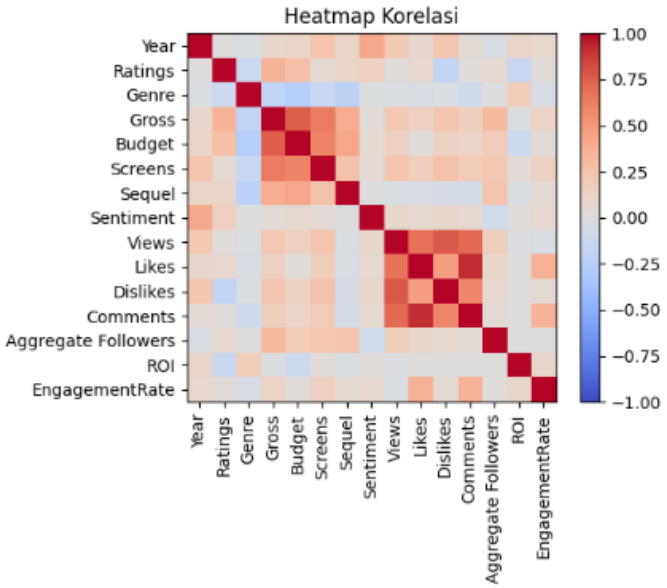
Pada tahap ini didapat hasil proporsi 0 sebesar 0,465, *unique values* 25, varians 33,166, korelasi ke Gross r = 0,036, dan korelasi ke ROI r = 0,017. Angka korelasi yang didapat lemah, tetapi proporsi 0 tidak mendominasi dan variasi masih ada, sehingga kolom Sentiment dipertahankan. Kolom ini digunakan sebagai indikator pendukung.

7. Penilaian kebutuhan PCA

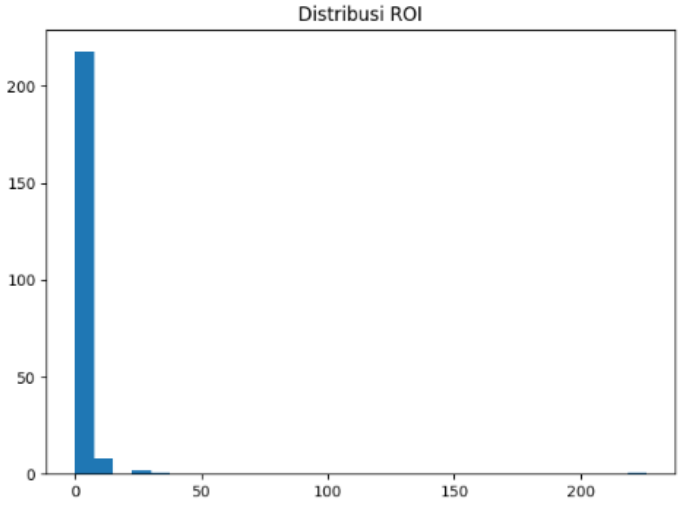
Tahap ini menilai apakah PCA (*Principal Component Analysis)* diperlukan untuk reduksi dimensi. Kriteria yang dipakai adalah jika korelasi maksimum antar metrik sosial ≥ 0,92, maka proses PCA dipertimbangkan untuk dijalankan. Hasil dari tahap ini adalah korelasi maksimum tercatat 0,917. Nilai tersebut masih berada di bawah ambang batas, maka proses PCA tidak perlu dilakukan.

8. Dataset final dan visual dasar

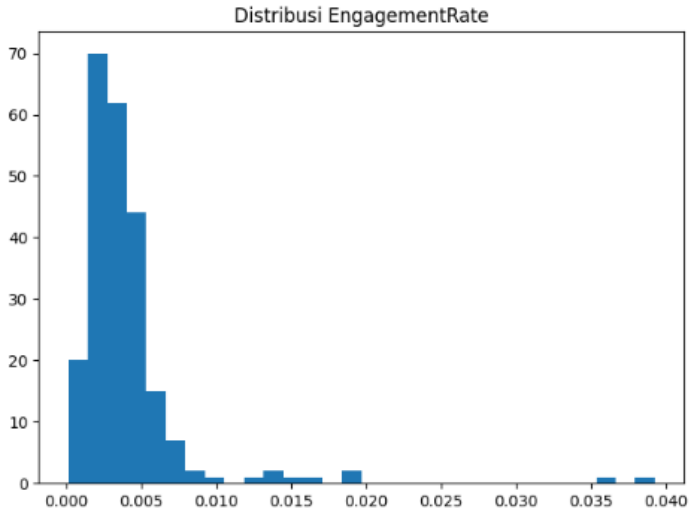
Hasil *preprocessing* disimpan sebagai file csm_final_minimal.csv. Hasil tidak memiliki nilai kosong dan data berisi 230 baris serta 16 kolom. Kolom yang dipakai, yaitu Movie, Year, Ratings, Genre, Gross, Budget, Screens, Sequel, Sentiment, Views, Likes, Dislikes, Comments, dan Aggregate Followers serta dua kolom turunan baru yaitu ROI dan EngagementRate.



Gambar 9. Heatmap korelasi



Gambar 10. Distribusi ROI



Gambar 11. Distribusi EngagementRate

KESIMPULAN

Proses *preprocessing* pada dataset CSM Movies 2014 dan 2015 menghasilkan data yang lebih bersih, konsisten, dan siap dianalisis. Langkah yang dijalankan meliputi validasi nilai, penanganan *missing values*, penambahan kolom turunan ROI dan EngagementRate, serta pengendalian outlier pada Gross dan Budget dengan winsorizing berbasis IQR. Dataset akhir bebas dari *missing values* dan struktur kolom tetap mudah diinterpretasi.

Setelah outlier dikendalikan, distribusi ROI menjadi lebih stabil. Budget berhubungan positif dengan Gross, sedangkan metrik sosial seperti Views dan Likes cenderung selaras dengan performa pendapatan. PCA tidak diperlukan karena tidak ada korelasi yang sangat tinggi antar metrik sosial. Dengan demikian, dataset hasil preprocessing sudah layak digunakan sebagai dasar analisis deskriptif.

REFERENSI

Ahmed, M. 2015. *CSM (Conventional and Social Media Movies) Dataset 2014 and 2015*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/424/csm+conventional+and+social+media+movies+dataset+2014+and+2015