

Analisis Data Eksploratif (EDA) pada Dataset Netflix dan TMDB untuk Memahami Faktor yang Mempengaruhi Performa Film dan Popularitas

Karina Auliasari¹, Febriana Santi Wahyuni², Ahmadun Siyamuladi³

^{1,2,3}Fakultas Teknologi Industri Institut Teknologi Nasional, Malang, 65143

¹karina.auliasari@lecturer.itn.ac.id, ², ³2318017@student.itn.ac.id



Histori Artikel:

Diajukan: 29 Oktober 2025

Disetujui: 25 June 2021

Dipublikasi: 30 June 2021

Kata Kunci:

Exploratory Data Analysis,
Data Science, Visualization,
Heatmap,
Netflix, Movies, Tmdb,

Digital Transformation

*Technology (Digitech) is an
Creative Commons License This
work is licensed under a
Creative Commons Attribution-
NonCommercial 4.0
International (CC BY-NC 4.0).*

Abstrak

Penelitian ini membahas penerapan Exploratory Data Analysis (EDA) pada dataset hiburan yang berisi informasi terkait film di platform Netflix dan data finansial dari TMDB. Tujuan dari penelitian ini adalah untuk memahami pola, tren, serta faktor-faktor yang berpotensi memengaruhi performa finansial dan popularitas film berdasarkan variabel seperti genre, durasi, rating, anggaran produksi, pendapatan, dan Return on Investment (ROI). Tahapan penelitian meliputi proses pembersihan data, penggabungan dua dataset dengan struktur kolom yang serupa, penambahan fitur baru seperti genre utama (main genre) untuk memperkaya analisis, serta visualisasi hasil eksplorasi. Hasil penelitian menunjukkan bahwa film dengan genre tertentu, durasi menengah hingga panjang, dan rating populer cenderung memiliki pendapatan dan ROI lebih tinggi. Selain itu, tren tahunan menunjukkan fluktuasi dalam produksi dan konsumsi film, serta variasi preferensi penonton terhadap genre. Penerapan EDA pada dataset ini memberikan gambaran komprehensif mengenai karakteristik film dan hubungan antarvariabel yang relevan dalam konteks analisis industri hiburan dan strategi pemasaran film.

PENDAHULUAN

Industri hiburan, khususnya melalui platform streaming seperti Netflix, telah mengalami transformasi digital yang masif dan mengubah lanskap konsumsi media secara global maupun di Indonesia. Pertumbuhan pesat ini tidak hanya mengubah cara konsumen mengakses konten tetapi juga menghasilkan volume data yang sangat besar dari berbagai platform digital. Data ini mencakup preferensi penonton, demografi pengguna, interaksi konten, kinerja finansial film, hingga tren pasar yang dinamis. Dalam lingkungan yang sangat kompetitif antar penyedia layanan streaming, pemanfaatan data menjadi krusial untuk pengambilan keputusan strategis, mulai dari produksi konten hingga strategi pemasaran digital yang efektif. Sejalan dengan temuan Athaya, dkk. (2024), "Prediksi Kesuksesan Film Indonesia" menjadi salah satu fokus utama, di mana data historis digunakan untuk memproyeksikan potensi performa film di masa depan. Perusahaan seperti Netflix, yang keputusan berlangganannya dipengaruhi berbagai faktor, serta basis data komprehensif seperti The Movie Database (TMDB), menjadi sumber data utama untuk menganalisis faktor-faktor seperti genre, aktor, dan ulasan penonton yang berkontribusi terhadap kesuksesan sebuah film.

Untuk mengekstrak wawasan berharga dari tumpukan data mentah tersebut, diperlukan sebuah pendekatan metodis sebelum melangkah ke pemodelan yang kompleks. Di sinilah Exploratory Data Analysis (EDA) atau Analisis Data Eksploratif memainkan peran vital. EDA adalah "proses investigasi awal pada data untuk menemukan pola, melihat anomali, menguji hipotesis, dan memeriksa asumsi" dengan bantuan statistik deskriptif dan representasi grafis (visualisasi). Sebagai sebuah pendekatan yang "bersifat terbuka dan fleksibel", EDA memungkinkan analis untuk menggali informasi atau permasalahan yang relatif masih baru dan "memahami sifat dan karakteristik data secara lebih mendalam". Melalui teknik seperti visualisasi data dan perhitungan statistik ringkasan, EDA membantu mengidentifikasi tren utama, distribusi data, hubungan antar variabel, serta potensi outlier atau bias dalam dataset. Tanpa EDA, analisis berisiko menarik kesimpulan yang salah atau membangun model prediksi yang tidak akurat karena belum memahami sepenuhnya konteks dan struktur data yang mendasarinya.

Penelitian ini bertujuan untuk menerapkan metode Exploratory Data Analysis (EDA) secara mendalam pada dua dataset kaya dari industri hiburan, yaitu katalog film Netflix dan data film dari The Movie Database (TMDB). Fokus utama analisis adalah untuk mengidentifikasi dan memahami faktor-faktor kunci—seperti genre, anggaran (budget), pemeran (cast), durasi, tanggal rilis, rating, dan popularitas—yang berkorelasi signifikan dengan performa finansial dan tingkat popularitas film di platform digital. Sebagaimana ditekankan oleh Wibowo (2025), "pengetahuan sebelumnya tentang keberhasilan atau kegagalan tertentu film dan faktor apa yang mempengaruhi kesuksesan akan menguntungkan rumah produksi" dalam pengambilan keputusan bisnis seperti

pra-produksi, strategi pemasaran, hingga pengembangan sistem rekomendasi film yang lebih personal. Dengan memanfaatkan teknik visualisasi dan statistik deskriptif, penelitian ini akan mengeksplorasi hubungan antar variabel untuk menjawab pertanyaan tentang mengapa beberapa film, termasuk yang hadir di Netflix Indonesia, lebih populer atau berkinerja lebih baik daripada yang lain di era streaming saat ini.

STUDI LITERATUR

Penelitian mengenai faktor-faktor yang mempengaruhi kesuksesan film telah banyak dilakukan, terutama dengan ketersediaan dataset publik berskala besar. Banyak peneliti memanfaatkan data ini untuk membangun model prediktif. Sebagai contoh, Wibowo (2025) dalam penelitiannya menggunakan dataset The Movie Database (TMDB) untuk memprediksi peringkat film menggunakan metode Random Forest Regression. Studi tersebut menekankan bahwa "pengetahuan sebelumnya tentang keberhasilan atau kegagalan tertentu film. akan menguntungkan rumah produksi," yang menyoroti nilai praktis dari analisis data dalam industri ini. Demikian pula, Athaya, dkk. (2024) memfokuskan analisis mereka pada "Prediksi Kesuksesan Film Indonesia" dengan menerapkan Jaringan Syaraf Tiruan, menunjukkan bahwa variabel seperti genre, sutradara, dan aktor memiliki pengaruh signifikan terhadap penerimaan pasar.

Selain pemodelan prediktif, Exploratory Data Analysis (EDA) diakui sebagai langkah fundamental dalam proses data mining untuk menggali wawasan awal. EDA berfokus pada penggunaan statistik deskriptif dan visualisasi untuk mengidentifikasi tren, mendeteksi anomali, dan memahami struktur data. Pratama & Sari (2023) mendefinisikan EDA sebagai metode krusial untuk "memahami sifat dan karakteristik data secara lebih mendalam." Dalam konteks dataset film, EDA memungkinkan peneliti untuk mengeksplorasi distribusi variabel seperti anggaran, pendapatan, dan durasi, serta menganalisis hubungan kategorikal antara genre dan popularitas sebelum melangkah ke pemodelan yang lebih kompleks.

Meskipun banyak penelitian yang berfokus pada akurasi prediksi, terdapat kebutuhan untuk studi yang secara khusus mendalami temuan dari proses EDA itu sendiri. Penelitian sebelumnya seringkali menggunakan EDA hanya sebagai langkah awal, namun tidak melaporkan secara komprehensif wawasan yang didapat dari eksplorasi tersebut. Studi ini berupaya mengisi celah tersebut dengan menjadikan EDA sebagai fokus utama. Dengan menganalisis dataset Netflix dan TMDB secara eksploratif, penelitian ini tidak hanya mengidentifikasi variabel yang berkorelasi dengan popularitas, tetapi juga memvisualisasikan pola-pola tersebut untuk memberikan pemahaman intuitif tentang apa yang mendorong performa film di era streaming.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *Exploratory Data Analysis (EDA)* untuk memahami pola, tren, dan distribusi data dalam dataset Netflix dan TMDB. Tahapan penelitian dilakukan secara sistematis mulai dari pengumpulan data, pembersihan data, penggabungan, hingga visualisasi dan interpretasi hasil.

1. Sumber Data

- Data pertama yang digunakan adalah `netflix_titles.csv` yang berisikan `show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description`.
- Data kedua yang digunakan adalah `tmdb_5000_movies.csv` yang berisikan `budget, genres, homepage, id, keywords, original_language, overview, popularity, production_company, production_countries, release_date, revenue, runtime, spoken_language, status, tagline, title, vote_average, vote_count`.

Kedua dataset ini digunakan untuk menganalisis dan membandingkan faktor-faktor risiko yang berkontribusi terhadap kedua penyakit tersebut.

2. Tahapan Analisis

a. Import Library

Proses ini melibatkan penggunaan pustaka *Python* seperti *pandas* untuk manipulasi data, *numpy* untuk komputasi numerik, *matplotlib* dan *seaborn* untuk visualisasi data. Semua analisis dilakukan di platform *Google Colab* karena mendukung integrasi langsung dengan *Google Drive*.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from google.colab import drive
drive.mount('/content/drive')
```

Gambar 1.1 Proses import library

b. Data Loading dan Pemeriksaan Awal

Dataset dimuat dari Google Drive menggunakan `pandas.read_csv()` dan diperiksa menggunakan fungsi `head()` untuk mengetahui struktur kolom, tipe data, dan jumlah nilai kosong.

```
df_netflix = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/netflix_titles.csv")
df_tmdb = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/tmdb_5000_movies.csv")

df_netflix.head()
```

Gambar 1.2.Proses load dataset lewat google drive

c. Data Cleansing (Pembersihan Data)

1) Penghapusan duplikat data berdasarkan title.

```
df_netflix.drop_duplicates(subset='title', inplace=True)
```

Gambar 1.3 Proses cleansing data yang duplikat

2) Normalisasi Menghilangkan whitespace pada teks.

```
for col in ['title', 'director', 'country', 'listed_in']:
    df_netflix[col] = df_netflix[col].astype(str).str.strip()
```

Gambar 1.4 Proses normalisasi menghilangkan whitespace

3) Normalisasi menangani missing values.

```
# Tangani missing values
df_netflix.loc[:, 'director'] = df_netflix['director'].fillna('Unknown')
df_netflix.loc[:, 'country'] = df_netflix['country'].fillna('Unknown')
df_netflix.loc[:, 'rating'] = df_netflix['rating'].fillna('Not Rated')
df_netflix.loc[:, 'duration'] = df_netflix['duration'].fillna('Unknown')
```

Gambar 1.5 Proses Menangani missing values

4) Filtering Movies.

```
netflix_movies = df_netflix[df_netflix['type'] == 'Movie'].copy()
```

Gambar 1.6 Proses Filtering Movies

5) Normalisasi kolom duration ke menit

```
netflix_movies['duration_min'] = (
    netflix_movies['duration']
    .str.replace(' min', '', regex=False)
    .replace('Unknown', np.nan)
    .astype(float)
)
```

Gambar 1.7 Proses Normalisasi Duration ke menit

6) Memilih kolom yang relevan

```
netflix_clean = netflix_movies[['
    'title', 'director', 'country', 'release_year',
    'rating', 'duration_min', 'listed_in']]
```

Gambar 1.8 Proses Memilih kolom yang relevan

7) Menghapus Duplikasi berdasarkan title

```
df_tmdb.drop_duplicates(subset='title', inplace=True)
```

Gambar 1.9 Proses Menghapus Duplikasi

8) Normalisasi Menangani missing values

```
df_tmdb.fillna({
    'budget': 0,
    'revenue': 0,
    'runtime': df_tmdb['runtime'].median()
}, inplace=True)
```

Gambar 1.10 Proses Normalisasi menangani missing values

9) Normalisasi Menghilangkan whitespace pada teks

```
df_tmdb['title'] = df_tmdb['title'].astype(str).str.strip()
```

Gambar 1.11 Proses Normalisasi Menghilangkan whitespace pada teks

10) Normalisasi kolom release date menjadi datetime dan menambah kolom release year

```
df_tmdb['release_date'] = pd.to_datetime(df_tmdb['release_date'], errors='coerce')
df_tmdb['release_year'] = df_tmdb['release_date'].dt.year
```

Gambar 1.12 Proses Normalisasi menambah kolom release year

- 11) Menambahkan Kolom Profit dan ROI

```
df_tmdb['profit'] = df_tmdb['revenue'] - df_tmdb['budget']
df_tmdb['roi'] = df_tmdb['profit'] / df_tmdb['budget']
df_tmdb['roi'] = df_tmdb['roi'].replace([np.inf, -np.inf], np.nan)
```

Gambar 1.13 Proses Normalisasi Menambahkan Kolom Profit dan ROI

- 12) Filter film yang mempunyai data keuangan yang valid

```
df_tmdb_clean = df_tmdb[(df_tmdb['budget'] > 0) & (df_tmdb['revenue'] > 0)]
```

Gambar 1.14 Proses Normalisasi Filter film

- 13) Memilih kolom yang relevan

```
tmdb_clean = df_tmdb[['title', 'budget', 'revenue', 'profit', 'roi',
                       'popularity', 'runtime', 'release_year', 'genres']]
```

Gambar 1.15 Proses Memilih kolom yang relevan

- d. Data Enrichment (Penggabungan Data)

Dataset Netflix digabung dengan dataset TMDB berdasarkan title dan tahun rilis.

```
merged = pd.merge(
    netflix_clean,
    tmdb_clean,
    on=['title', 'release_year'],
    how='inner'
)
```

Gambar 1.16 Proses enrichment data dengan dataset kedua

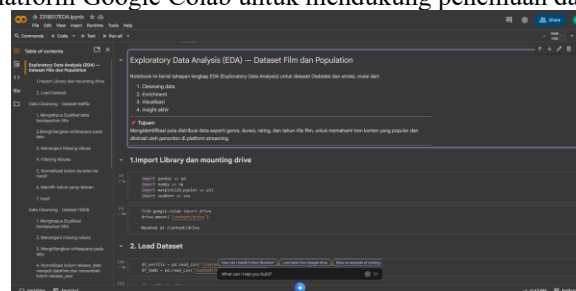
- e. Visualisasi Data

Data divisualisasikan menggunakan grafik batang (*bar chart*), histogram, boxplot, scatter plot, dan heatmap untuk menggambarkan:

- Distribusi kategori
- Distribusi budget dan revenue
- Korelasi numerik.
- Tren jumlah film per tahun.
- Rata – rata revenue per tahun.
- Distribusi ROI
- Top 10 film paling menguntungkan
- Durasi vs revenue
- Top movies berdasarkan revenue

HASIL

Bagian ini membahas hasil penerapan Exploratory Data Analysis (EDA) pada dataset Netflix dan TMDB yang telah melalui tahap pembersihan dan penggabungan. Analisis dilakukan untuk memahami karakteristik, pola, dan hubungan antar variabel dalam dataset. Seluruh visualisasi dibuat menggunakan pustaka Python seperti Matplotlib dan Seaborn di platform Google Colab untuk mendukung penemuan dan interpretasi hasil.



Gambar 1.17 Platform Google Colab

PEMBAHASAN

1. Statistik Deskriptif

Analisis statistik deskriptif dilakukan untuk memberikan gambaran umum tentang data numerik,

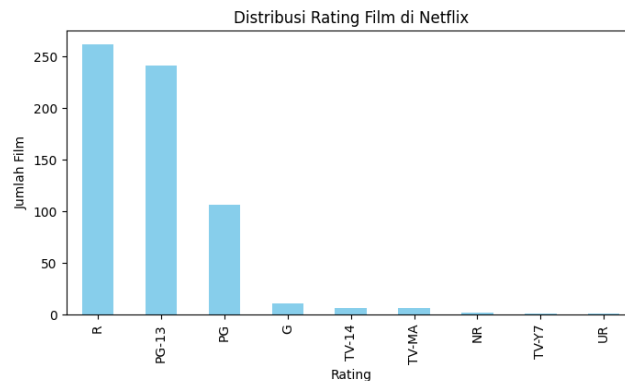
```
merged[['budget', 'revenue', 'profit', 'roi', 'runtime', 'duration_min', 'popularity']].describe()
```

	budget	revenue	profit	roi	runtime	duration_min	popularity
count	6.360000e+02	6.360000e+02	6.360000e+02	579.000000	636.000000	636.000000	636.000000
mean	4.241258e+07	1.235326e+08	8.112006e+07	26.330540	110.919811	111.301887	30.441629
std	4.473448e+07	1.546302e+08	1.281102e+08	535.675617	21.395228	21.423000	22.918237
min	0.000000e+00	0.000000e+00	-8.500000e+07	-1.000000	53.000000	45.000000	0.043023
25%	1.000000e+07	2.276873e+07	3.283713e+06	0.449716	96.000000	96.000000	14.812286
50%	2.700000e+07	7.056890e+07	3.582120e+07	1.545362	107.000000	107.500000	24.605541
75%	6.200000e+07	1.686150e+08	1.104288e+08	3.457436	121.000000	122.000000	40.245573
max	2.700000e+08	1.118889e+09	1.024889e+09	12889.386667	229.000000	229.000000	167.583710

Gambar 1.18 Proses Statistik Deskriptif

Dataset terdiri dari 636 film dengan variasi anggaran produksi yang luas, rata-rata sekitar 42 juta, dan beberapa film memiliki anggaran hingga 270 juta. Pendapatan film sangat bervariasi, dengan rata-rata 123 juta dan puncak lebih dari 1 miliar, sementara laba rata-rata mencapai 81 juta, meski ada film yang merugi. ROI menunjukkan keuntungan rata-rata 26 kali lipat, namun terdapat nilai ekstrem baik positif maupun negatif. Durasi film rata-rata sekitar 110 menit, dan popularitas film juga beragam dengan nilai rata-rata 30, mencerminkan perbedaan penerimaan penonton terhadap film-film tersebut.

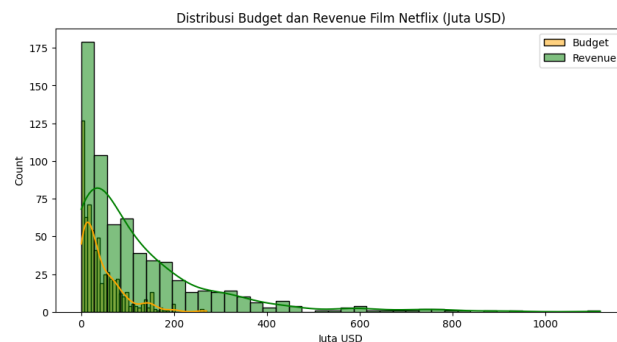
2. Distribusi kategori



Gambar 1.19 Tahapan visualisasi Distribusi kategori

Distribusi rating film di Netflix menunjukkan bahwa mayoritas film memiliki rating R dan PG-13, dengan masing-masing sekitar 260 dan 240 film. Rating PG juga cukup banyak dengan lebih dari 100 film. Sementara itu, film dengan rating yang lebih rendah seperti G, TV-14, dan TV-MA relatif sedikit, dan film dengan rating sangat rendah atau khusus seperti NR, TV-Y7, dan UR hampir tidak ada. Hal ini mengindikasikan bahwa Netflix lebih banyak menyediakan konten dengan rating yang lebih dewasa dan remaja, sementara konten untuk anak-anak atau tanpa rating cenderung lebih terbatas.

3. Distribusi budget dan revenue

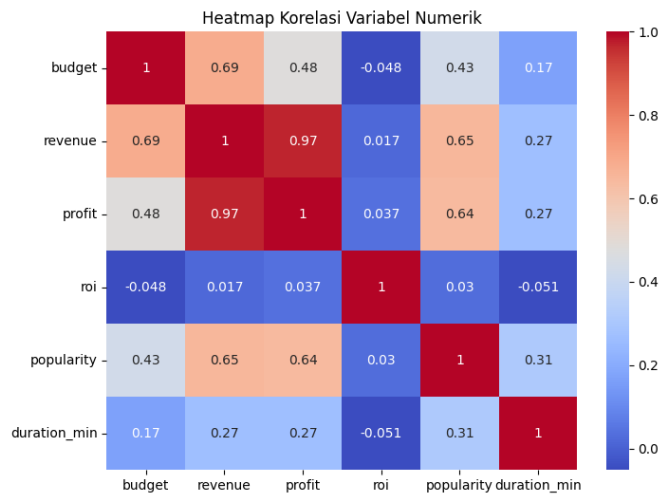


Gambar 1.20 Tahapan visualisasi data distribusi budget dan revenue

Distribusi anggaran (budget) dan pendapatan (revenue) film Netflix menunjukkan bahwa sebagian besar film memiliki budget di bawah 50 juta USD, dengan puncak distribusi yang cukup tajam pada kisaran tersebut. Sebaliknya, pendapatan film tersebar lebih luas dengan jumlah film yang menghasilkan revenue rendah hingga sangat tinggi, bahkan ada film dengan pendapatan lebih dari 1000 juta USD. Grafik ini

menunjukkan bahwa walaupun banyak film dibuat dengan anggaran relatif kecil, beberapa film berhasil meraih pendapatan yang jauh lebih besar, mengindikasikan adanya variasi besar dalam tingkat kesuksesan finansial film-film Netflix..

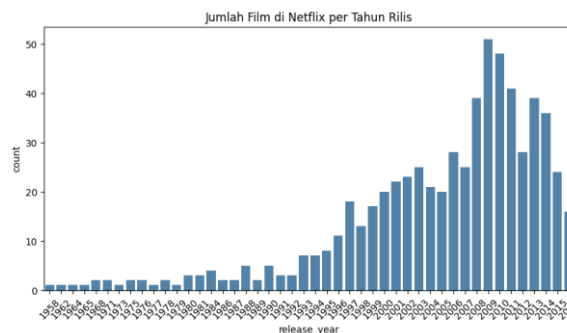
4. Korelasi numerik antar fitur utama



Gambar 1.21 Tahapan visualisasi data perbandingan fitur numerik

Hubungan yang paling signifikan terdapat antara revenue dan profit, sedangkan ROI dan duration_min relatif independen terhadap variabel lain. Ini menunjukkan bahwa faktor utama yang saling berkaitan dalam konteks ini adalah anggaran, pendapatan, keuntungan, dan popularitas.

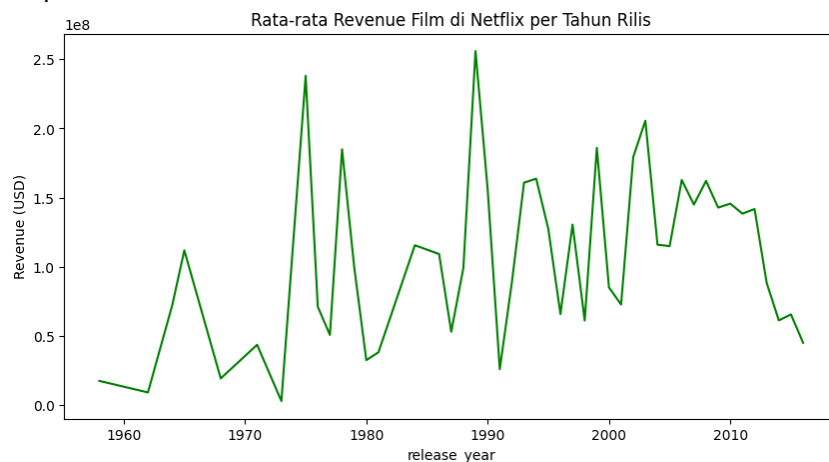
5. Tren jumlah film Netflix per tahun



Gambar 1.22 Tahapan visualisasi data tren film Netflix per tahun

Grafik menunjukkan bahwa produksi film yang masuk ke katalog Netflix meningkat pesat pada dua dekade terakhir, terutama di kisaran tahun 2000–2010, yang bisa jadi merupakan masa ekspansi besar Netflix terhadap konten film. Namun, setelah itu terjadi penurunan jumlah film baru, kemungkinan karena pergeseran fokus ke produksi konten orisinal atau kebijakan kurasi yang lebih selektif.

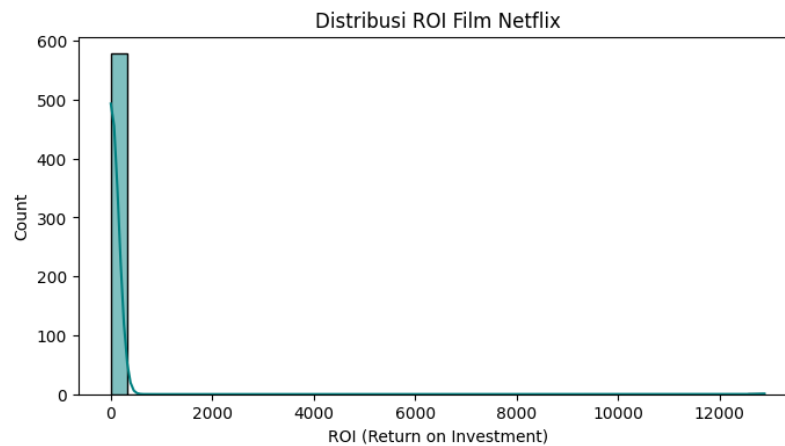
6. Rata-rata revenue per tahun



Gambar 1.23 Tahapan visualisasi data untuk rata – rata revenue

peningkatan rata-rata revenue film di Netflix yang sangat signifikan dan konsisten seiring dengan bertambahnya tahun rilis, di mana revenue film dari era klasik (1960-1990) berada pada level yang relatif rendah dan stabil, kemudian mengalami lonjakan spektakuler mulai tahun 2000-an dan terus melesat hingga era modern, mengindikasikan bahwa konten dengan tahun rilis lebih baru secara finansial jauh lebih dominan dan sukses di platform streaming ini, didorong oleh perubahan preferensi audiens dan pertumbuhan pasar digital.

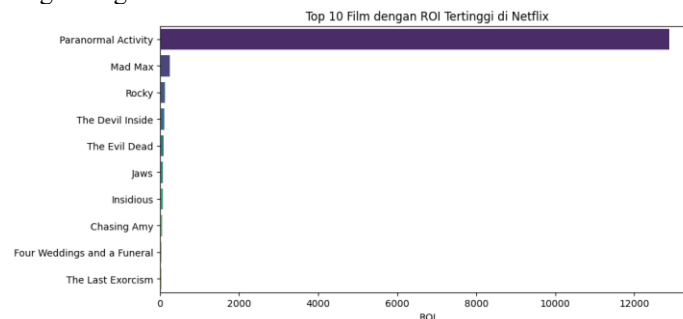
7. Distribusi ROI



Gambar 1.24 Tahapan visualisasi data distribusi ROI

Sebagian besar film memiliki Return on Investment yang rendah, dengan distribusi yang sangat miring ke kanan (right-skewed) di mana mayoritas film terkonsentrasi pada kisaran ROI 0 hingga 2000, sementara hanya segelintir film yang berhasil mencapai ROI sangat tinggi hingga di atas 4000, mengindikasikan bahwa kesuksesan finansial yang luar biasa hanya dialami oleh sebagian kecil konten, sedangkan sebagian besar film menghasilkan pengembalian investasi yang terbatas.

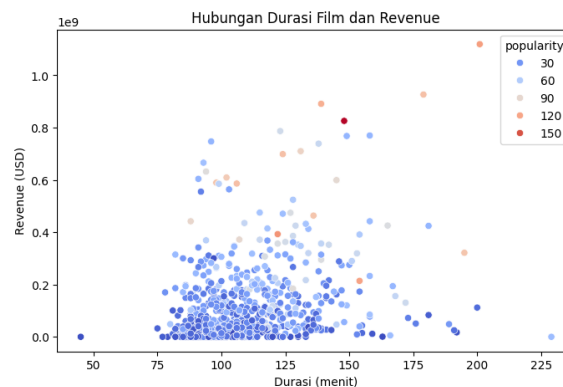
8. Top 10 Film Paling Menguntungkan



Gambar 1.25 Tahapan visualisasi data top 10 film

Mayoritas film dalam daftar 10 besar ini berasal dari genre horor dan thriller yang mendominasi peringkat, menunjukkan bahwa genre tersebut cenderung memberikan pengembalian investasi yang sangat optimal dibandingkan genre lain di platform streaming Netflix.

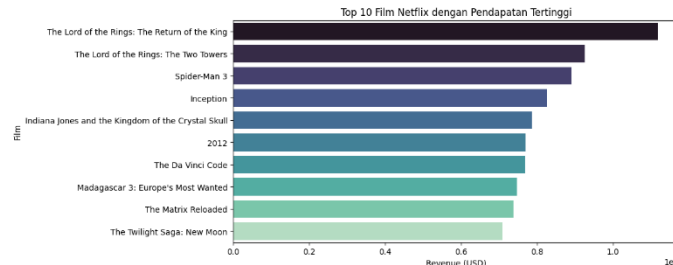
9. Durasi vs Revenue



Gambar 1.26 Tahapan visualisasi data durasi vs revenue

Berdasarkan grafik sebaran yang dianalisis, dapat disimpulkan bahwa tidak terdapat hubungan atau korelasi yang jelas antara durasi film dengan revenue yang dihasilkan, di mana film-film dengan berbagai variasi durasi—baik yang pendek (sekitar 60-90 menit) maupun yang panjang (sekitar 120-150 menit) menunjukkan sebaran revenue yang sangat acak tanpa pola tertentu, sehingga membuktikan bahwa durasi bukan merupakan faktor penentu utama terhadap kesuksesan finansial sebuah film di platform Netflix.

10. Top Movies Berdasarkan Revenue



Gambar 1.27 Tahapan visualisasi data top movies revenue

Berdasarkan grafik yang ditampilkan, dapat disimpulkan bahwa film "The Lord of the Rings: The Return of the King" menghasilkan revenue tertinggi yang mendekati angka maksimal, diikuti oleh sekuelnya "The Lord of the Rings: The Two Towers" dan "Spider-Man 2", sementara film-film seperti "The Twilight Saga: New Moon" dan "The Da Vinci Code" berada di peringkat terbawah dalam daftar ini, menunjukkan bahwa franchise fantasi dan superhero mendominasi peringkat atas film dengan pendapatan tertinggi.

11. Ringkasan/Insight EDA

Secara keseluruhan, hasil *Exploratory Data Analysis* memberikan beberapa temuan penting sebagai berikut :

1. Penggabungan Data: Dataset akhir terdiri dari 636 film yang memiliki data lengkap dari katalog Netflix dan detail finansial dari TMDB, setelah melalui proses pembersihan dan penggabungan.
2. Distribusi Finansial: Variabel budget, revenue, dan profit sangat condong ke kanan, menunjukkan bahwa sebagian besar film memiliki angka finansial yang relatif rendah, namun ada beberapa film blockbuster yang nilainya sangat tinggi.
3. ROI (Return on Investment): Median ROI sekitar 1.55, menandakan 50% film menghasilkan keuntungan minimal 55% dari anggarannya. ROI tidak berkorelasi kuat dengan budget atau popularity, artinya film berbudget rendah bisa sangat menguntungkan secara persentase.
4. Popularitas: Distribusi popularity juga condong ke kanan, dengan beberapa film yang jauh lebih populer daripada mayoritas lainnya. Popularitas berkorelasi positif cukup kuat dengan revenue dan profit.
5. Tren Tahunan: Jumlah film dalam dataset meningkat di tahun-tahun yang lebih baru (puncak sekitar 2016-2018), namun rata-rata revenue per tahun tidak menunjukkan tren naik/turun yang jelas.
6. Performa Film: Film dengan ROI tertinggi seringkali adalah film berbudget rendah yang sukses ("The Gallows"), sedangkan film dengan revenue tertinggi adalah blockbuster berbudget besar ("Avengers: Age of Ultron").

KESIMPULAN

Berdasarkan hasil *Exploratory Data Analysis* (EDA) terhadap dataset gabungan Netflix dan TMDB, dapat disimpulkan bahwa analisis eksploratif mampu memberikan pemahaman awal terhadap faktor-faktor yang mempengaruhi performa dan popularitas film sebelum dilakukan pemodelan atau analisis lanjutan. Hasil analisis menunjukkan distribusi data finansial (anggaran, pendapatan, keuntungan) yang sangat condong ke kanan, mengindikasikan dominasi oleh sejumlah kecil film blockbuster. Dari sisi metrik kesuksesan, popularitas menunjukkan korelasi positif yang cukup kuat dengan pendapatan dan keuntungan absolut. Namun, Return on Investment (ROI) memiliki korelasi yang sangat lemah dengan anggaran maupun popularitas, menyoroti bahwa film berbudget rendah pun dapat menghasilkan persentase pengembalian investasi yang sangat tinggi. Karakteristik lain seperti durasi film, meskipun cenderung terdistribusi normal (sekitar 111 menit), hanya memiliki korelasi lemah dengan metrik kesuksesan finansial. Secara keseluruhan, *Exploratory Data Analysis* efektif dalam mengidentifikasi hubungan kunci (seperti popularitas vs. profit), karakteristik data (distribusi skewed), dan nuansa penting (perilaku ROI yang berbeda) yang krusial untuk penelitian lanjutan, misalnya dalam membangun model prediktif kesuksesan film di platform streaming.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak yang telah mendukung penyelesaian penelitian ini, khususnya kepada Dosen Pengampu Mata Kuliah Ibu Karina Aulia Sari, ST. M.Eng dan Ibu Febriana Santi Wahyuni, S.Kom, M.Kom dan teman-teman yang memberikan masukan dalam proses analisis data. Penulis juga menyampaikan apresiasi kepada penyedia dataset Netflix dan TMDB di Kaggle yang telah memungkinkan pelaksanaan penelitian ini secara terbuka dan transparan.

REFERENSI

- Athaya, H., Amrulloh, I. T. A., & Mufidah, R. (2024). Prediksi Kesuksesan Film Indonesia Menggunakan Jaringan Syaraf Tiruan. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 2106–2111.
- Pratama, R. A., & Sari, D. P. (2023). Implementasi Exploratory Data Analysis untuk Analisis dan Visualisasi Data. *Jurnal Ilmiah Teknik Elektro (JITET)*, 12(2), 115-122.
- Wibowo, A. T. (2025). Analysis of The Movie Database Film Rating Prediction with Ensemble Learning Using Random Forest Regression Method. *Jurnal Teknik Informatika (JUTIF)*, 6(1), 1-10.
- Adawiyah, R., Yanti, S. M., Fadillah, M. N., & Kurnia, R. (2024). Analisis Kritik Penonton terhadap Film Adaptasi A Business Proposal Versi Indonesia: Perspektif Media Sosial. *Jurnal Riset Rumpun Agama dan Filsafat*, 3(2), 56-62
- Arief, I., & dkk. (2024). Analisis Popularitas Genre Film di Indonesia Menggunakan Algoritma Decision Tree. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(4), 5555–5563.
- Seso, A. C., & Rahmi, N. (2021). Analisis Efektivitas Kebijakan Pajak Hiburan Berbasis E-POS Dalam Meningkatkan Penerimaan Pajak di Masa Pandemi Covid-19 Pada Badan Pendapatan Daerah DKI Jakarta Tahun 2020. *Jurnal Ilmiah Mahasiswa Aplikasi Ilmu Administrasi Publik (JUMAIP)*, 1(3), 212-223.
- Bahranadinata, E. (2025). Penelitian Eksploratif: Pengertian, Tujuan, Ciri, dan Contohnya. *Manajemen Sistem Informasi Telkom University*.
- Pratama, R. A., & Sari, D. P. (2023). *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 12(2), 115-122. (Catatan: Sumber ini sama dengan nomor 2, namun diintegrasikan untuk memperkuat definisi EDA).
- Lestari, S., & Ramdhani, M. M. (2023). Sistem Rekomendasi Film Menggunakan Metode Content-Based Filtering Studi Kasus Materi Data Mining di SMK IDN Boarding School. *Jurnal Indonesia : Manajemen Informatika dan Komunikasi (JIMIK)*, 4(3), 1581–1587.
- Suwandy, N. I. F., & Fathonah, R. N. S. (2025). Sistem Rekomendasi TV Series Berdasarkan Genre Menggunakan Algoritma KNN. *Insologi: Jurnal Sains dan Teknologi*, 4(3), 273-282.
- Utami, D. B. (2019). Mengenal Indonesia Melalui Netflix Original Movie. *Jurnal Komunikasi*, 11(1), 70–80.
- Mudjiyanto, B. (t.t.). Tipe Penelitian Eksploratif Komunikasi. *Media Neliti*.
- Karolina, C. M., Maryani, E., & Sjchro, D. W. (2020). Implikasi genre film dan pemahaman penonton film tuna netra di 'Bioskop Harewos'. *ProTVF*, 4(1), 123.
- Laras, D., & Hasrullah, H. (Tahun?). Analisis Kinerja Sistem Rekomendasi Film Berbasis Deep Learning Menggunakan Model Neural Network Pada Dataset Movielens. *Jurnal Locus Penelitian dan Pengabdian*.
- Sitorus, T. A., & Astrianty, L. E. (Tahun?). Implementasi The Movie DataBase API Untuk Sistem Informasi Film Berbasis Mobile. *JUTISI: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*.