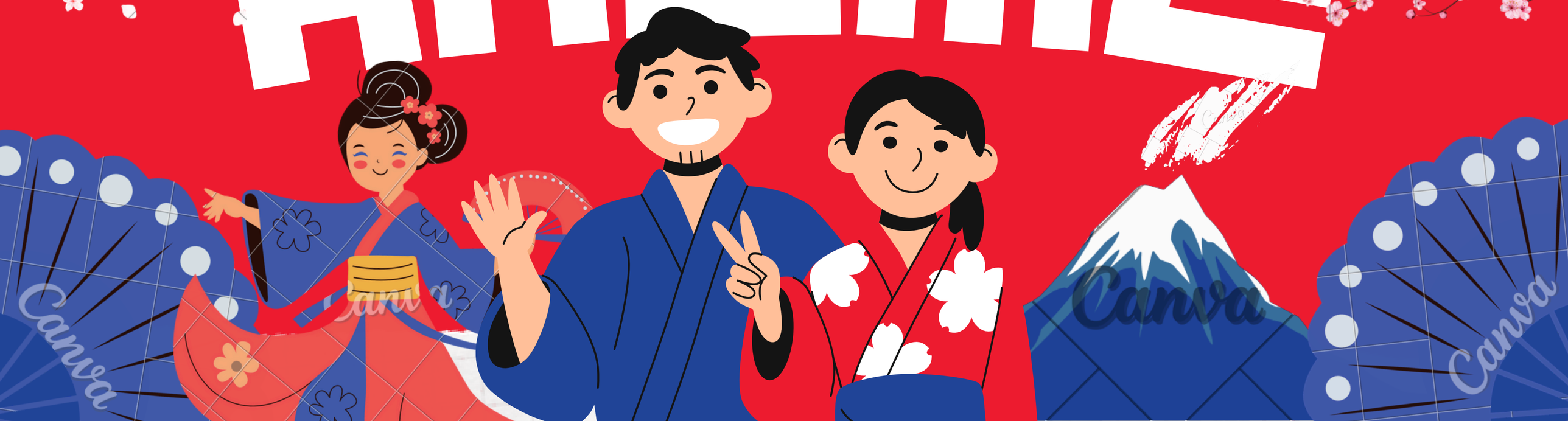
SILICON BALPEN

ANIME



KASDAD - SILICON BALPEN

OUTLINE

1. Introduction
2. Data Explanation
3. Data Reading
4. Exploratory Data Analysis
5. Data Preprocessing
6. Classification Modelling
7. Regression Modelling
8. Clustering Modelling

KASDAD - SILICON BALPEN

MEET THE TEAM

Caesar justitio
2206082373

Darrel Jeremiah
2206829225

Harjuno Abdullah
2206814053

Rayhan Dwi Sakha
2206082676

KASDAD - SILICON BALPEN

INTRODUCTION

Studi mengenai data anime ini bertujuan untuk memberikan pemahaman yang lebih mendalam mengenai berbagai aspek yang memengaruhi **performa dan karakteristik sebuah anime.** Analisis dilakukan melalui:

- **Klasifikasi** untuk memprediksi rating usia penonton berdasarkan fitur konten dan metadata,
- **Regresi** untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap jumlah penjualan, serta
- **Clustering** untuk menemukan pola dan segmentasi tersembunyi dalam dataset tanpa menggunakan label.

Pendekatan ini bertujuan memberikan wawasan yang berguna bagi industri anime dalam strategi produksi, distribusi, dan segmentasi audiens.

KASDAD - SILICON BALPEN

ABOUT DATASET

Anime Dataset berisi data dari 6.000 judul anime yang mencakup informasi seperti skor pengguna, penjualan, popularitas, rating usia, serta detail produksi.

Judul
Judul anime

Skor
Skor pengguna dari 1–10

Jumlah Penjualan
Total penjualan (DVD/digital, unit)

Peringkat
Peringkat berdasarkan skor.

Premier
Musim dan tahun rilis perdana

Popularitas
Peringkat berdasarkan popularitas.

Waktu Penayangan
Tanggal mulai dan selesai tayang

KASDAD - SILICON BALPEN

Status
Status penayangan (Tamat/Berjalan)

Durasi
Durasi tiap episode (menit)

Sumber
Asal materi
(Manga, Novel, dll.)

Episode
Total jumlah episode

Pemegang Lisensi
Perusahaan lisensi distribusi.

Studio
Studio animasi yang memproduksi anime

Produser
Nama perusahaan produksi

Rating
Klasifikasi usia penonton
(G, PG, dll.)

KASDAD - SILICON BALPEN

DATA READING

Tipe Data Fitur

Column	Dtype
id	int64
Judul	object
Skor	float64
Jumlah Penjualan	int64
Peringkat	int64
Popularitas	int64
Episode	object
Status	object
Waktu Penayangan	object
Premier	object
Produser	object
Pemegang Lisensi	object
Studio	object
Sumber	object
Durasi	object
Rating	object

Cek Missing Value

=== Missing Values ===	
id	0
Judul	0
Skor	0
Jumlah Penjualan	0
Peringkat	0
Popularitas	0
Episode	0
Status	0
Waktu Penayangan	0
Premier	3090
Produser	0
Pemegang Lisensi	0
Studio	0
Sumber	0
Durasi	0
Rating	20

Banyak Data

```
=== Informasi Data ===
Jumlah data: 6000
Jumlah atribut: 16
```

Data Duplikat

```
=== Data Duplikat ===

Jumlah data duplikat: 0
```

Data Outlier

```
Column: id, Outlier Count: 0, Outlier Percentage: 0.00%
Column: Skor, Outlier Count: 113, Outlier Percentage: 1.89%
Column: Jumlah Penjualan, Outlier Count: 722, Outlier Percentage: 12.07%
Column: Peringkat, Outlier Count: 0, Outlier Percentage: 0.00%
Column: Popularitas, Outlier Count: 94, Outlier Percentage: 1.57%
Column: Episode, Outlier Count: 468, Outlier Percentage: 7.83%
Column: Tahun_Tayang, Outlier Count: 217, Outlier Percentage: 3.63%
Column: Durasi_menit, Outlier Count: 1865, Outlier Percentage: 31.19%
```

KASDAD - SILICON BALPEN

DATA READING

Data Duplikat

```
judul_terduplikat = data["Judul"].value_counts()
judul_terduplikat = judul_terduplikat[judul_terduplikat > 1]
print(judul_terduplikat)
```

```
Judul
Emo FazeLaw of Devil                                                                                    3
Mato Seihei no SlaveChained Soldier                                                                     3
Shokugeki no Souma: San no Sara - Tootsuki Ressha-henFood Wars! The Third Plate: Totsuki Train Arc      2
Doraemon Movie 38: Nobita no TakarajimaDoraemon the Movie 2018: Nobita's Treasure Island                2
Senpai ga Uzai Kouhai no HanashiMy Senpai is Annoying                                                   2
                                                                                                       ..
Zettai Muteki Raijin-OhMatchless Raijin-Oh                                                              2
Mashiro-iro Symphony: The Color of LoversMashiroiro Symphony: The Color of Lovers                       2
Dungeon ni Deai wo Motomeru no wa Machigatteiru Darou ka IV: Fuka Shou - Yakusai-henIs It Wrong to Try to Pick Up Girls in a Dungeon? IV Part 2    2
Amagami SS+ Plus                                                                                        2
Mirai NikkiThe Future Diary OVA                                                                         2
```

Terdapat beberapa data duplikat yang tidak terdeteksi karena memiliki id yang berbeda

KASDAD - SILICON BALPEN

EXPLORATORY DATA ANALYSIS

Mencari *insight* untuk mendapatkan gambaran umum dari dataset ACD

Bagaimana Persebaran rating dari data tersebut? Mengapa rating itu yang paling banyak didapati di data tersebut?

Di musim apakah penjualan anime terendah terjadi? Berikan analisis kalian kenapa bisa terjadi hal ini!

Berapa jumlah anime yang tayang tiap tahun? Berikan minimal 2 anomali dari hasil observasi tersebut!

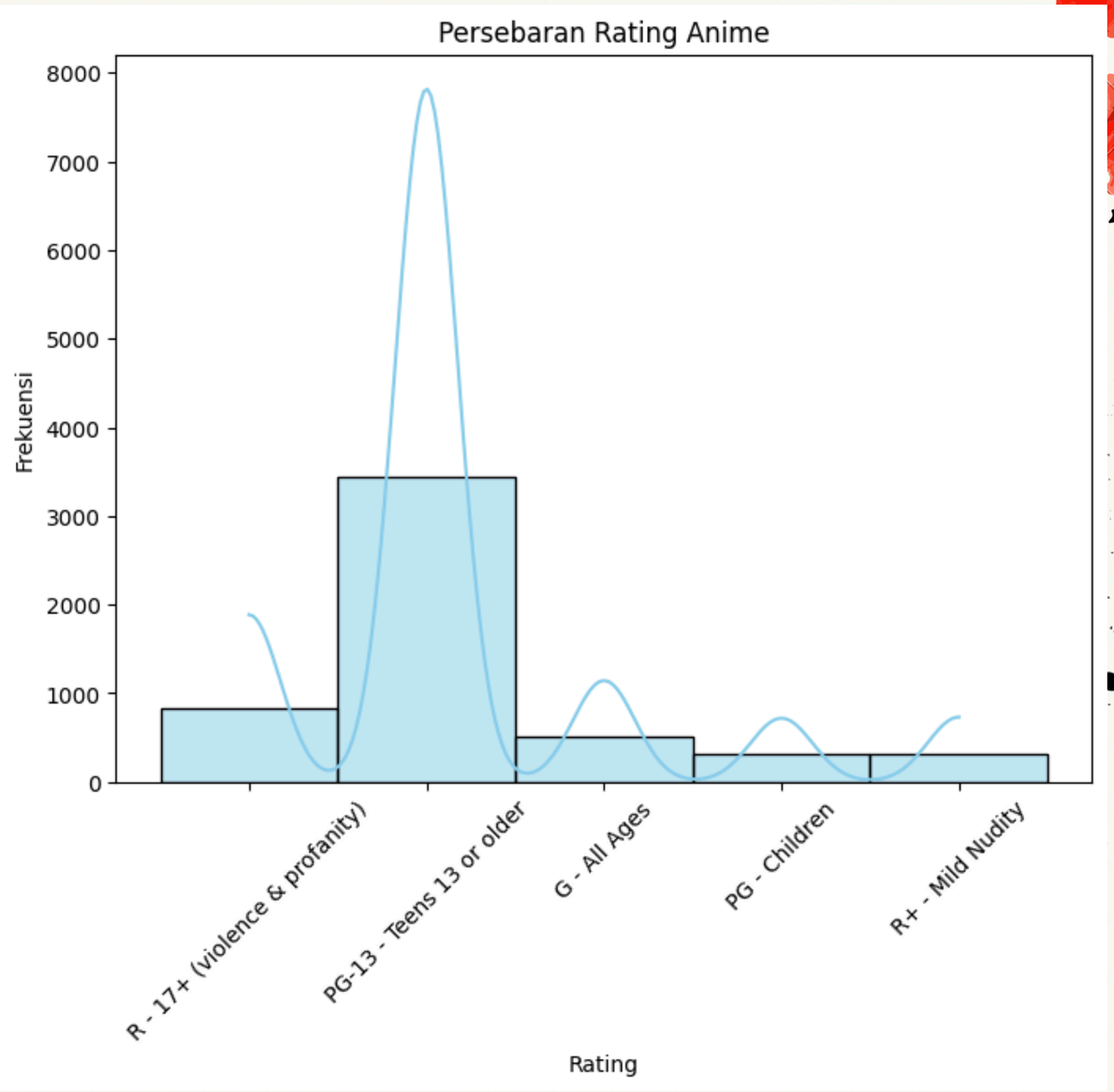
Produser apa yang menghasilkan penjualan anime terbanyak dari data tersebut? Mengapa produser tersebut yang menjadi penjual tertinggi?

Bagaimana hubungan antara Jumlah Penjualan dan Popularitas dengan Rating Anime?

Apakah anime dengan status berjalan dan status memiliki pengaruh terhadap skor dan penjualan anime tersebut?

KASDAD - SILICON BALPEN

1. Bagaimana Persebaran rating dari data tersebut? Mengapa rating itu yang paling banyak didapati di data tersebut?

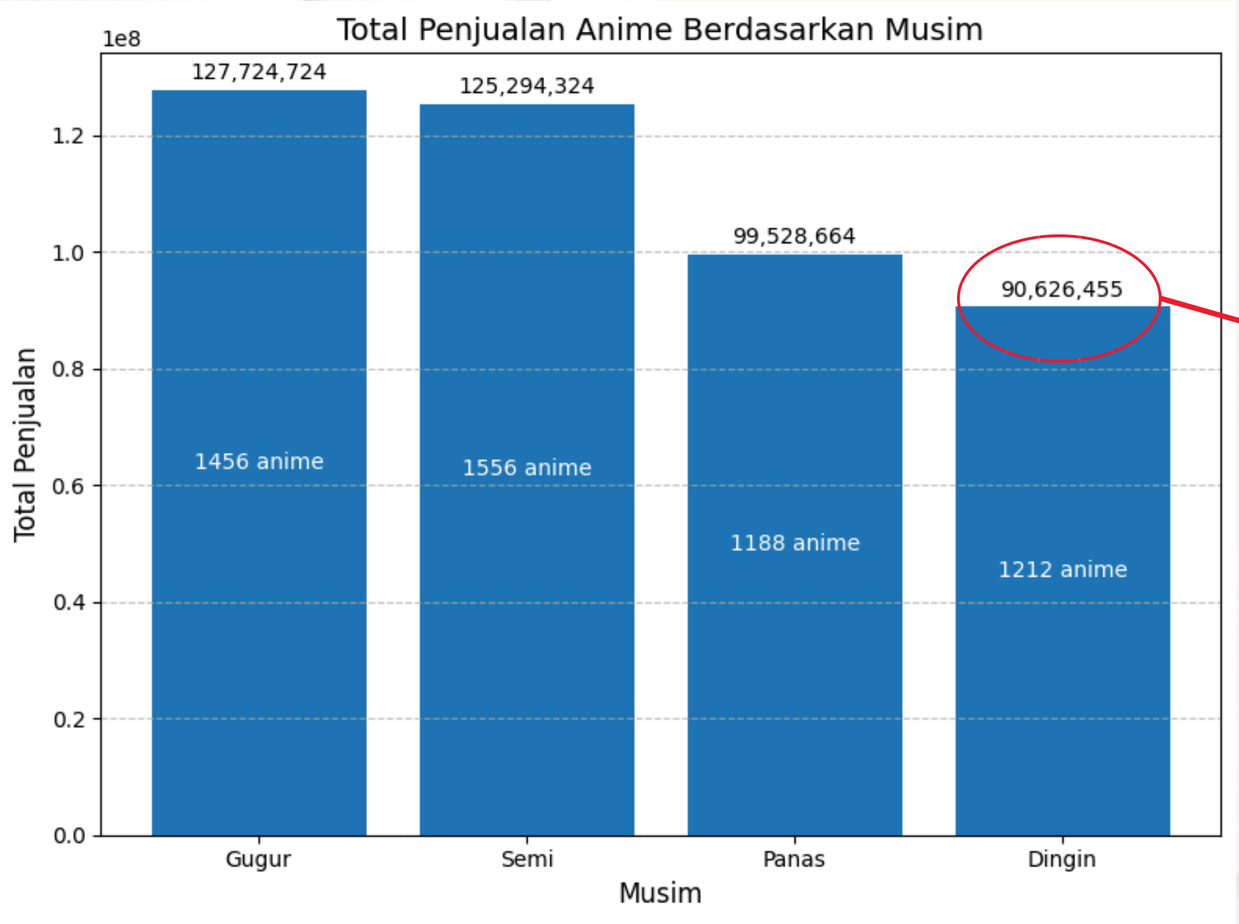


KASDAD - SILICON BALPEN

ANALISIS

1. Dominasi Rating “PG-13 – Teens 13 or older”
   - **63% anime dalam dataset memiliki rating PG-13.**
   - Menunjukkan fokus industri anime pada segmen remaja usia 13+ sebagai segmen pasar utama.
2. Distribusi Tidak Merata antar Kategori
   - R - 17+ menyusul PG-13 dalam jumlah, menunjukkan konten dewasa cukup umum.
   - Rating G - All Ages, PG - Children, dan R+ - Mild Nudity jauh lebih sedikit → **segmen anak-anak kurang dominan.**
3. Kecenderungan ke Rating Dewasa
   - **PG-13 dan R-17+ mendominasi**, mencerminkan banyaknya anime dengan tema kekerasan, kompleksitas psikologis, atau konten berat lainnya.
   - Anime ramah untuk anak lebih **terbatas** secara kuantitas.

2. Di musim apakah penjualan anime terendah terjadi? Berikan analisis kalian kenapa bisa terjadi hal ini!



Dingin

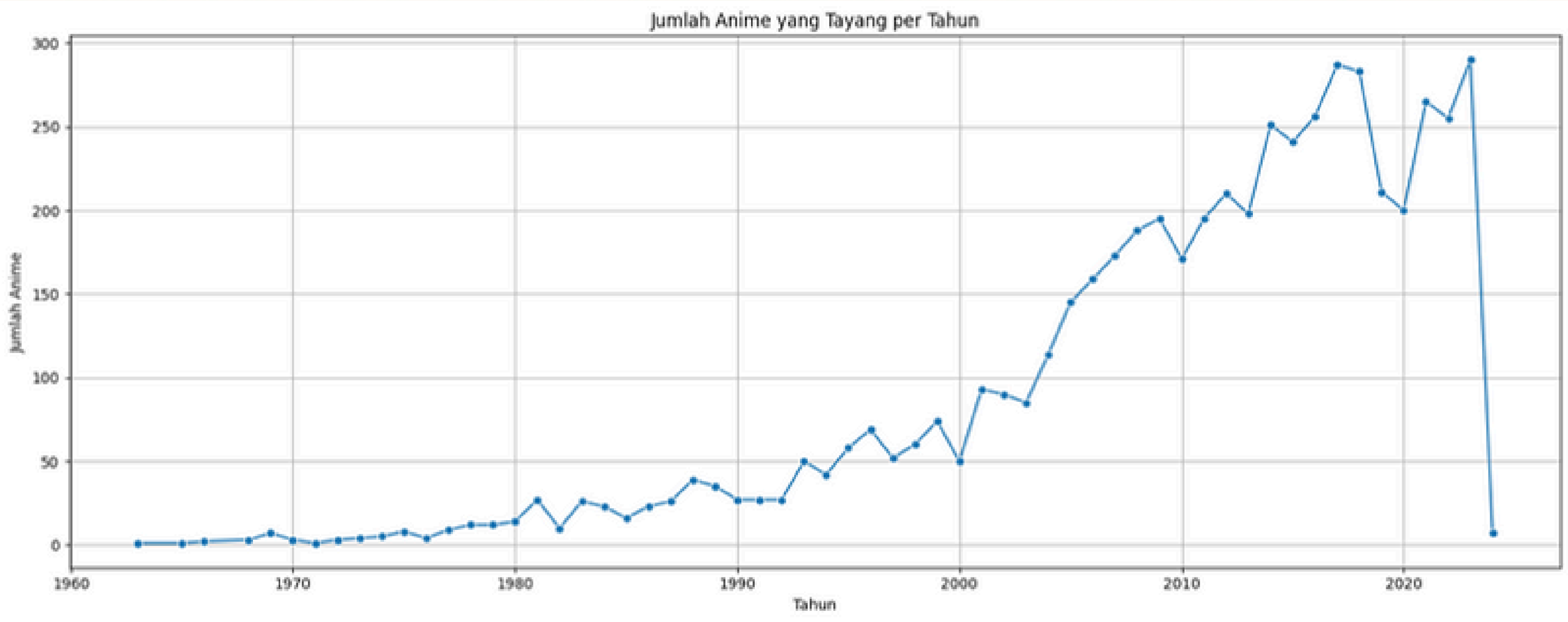
KASDAD - SILICON BALPEN

ANALISIS

Musim	Jumlah Penjualan	Skor	Popularitas	Episode	Jumlah Anime	Penjualan Rata-rata
Gugur	127724724	7.416868	3829.961538	18.926796	1456	87723.02
Panas	99528664	7.395842	3849.609428	10.284992	1188	83778.34
Semi	125294324	7.396060	4156.606684	20.281290	1556	80523.34
Dingin	90626455	7.405553	4104.886139	12.676936	1212	74774.30

1 .Siklus Industri
Musim dingin merupakan awal tahun dimana banyak studio masih dalam masa transisi yang mengakibatkan banyak rilisin besar ditahan hingga musim semi atau gugur.
2. Perilaku Konsumen
Siklus industri tadi juga dipengaruhi karena perilaku konsumen. Banyak Pengeluaran selain Anime pada musim ini, seperti Natal dan Tahun Baru. Jadi pada periode ini fokus utama konsumen bukan ke pembelian Anime

3. Berapa jumlah anime yang tayang tiap tahun?



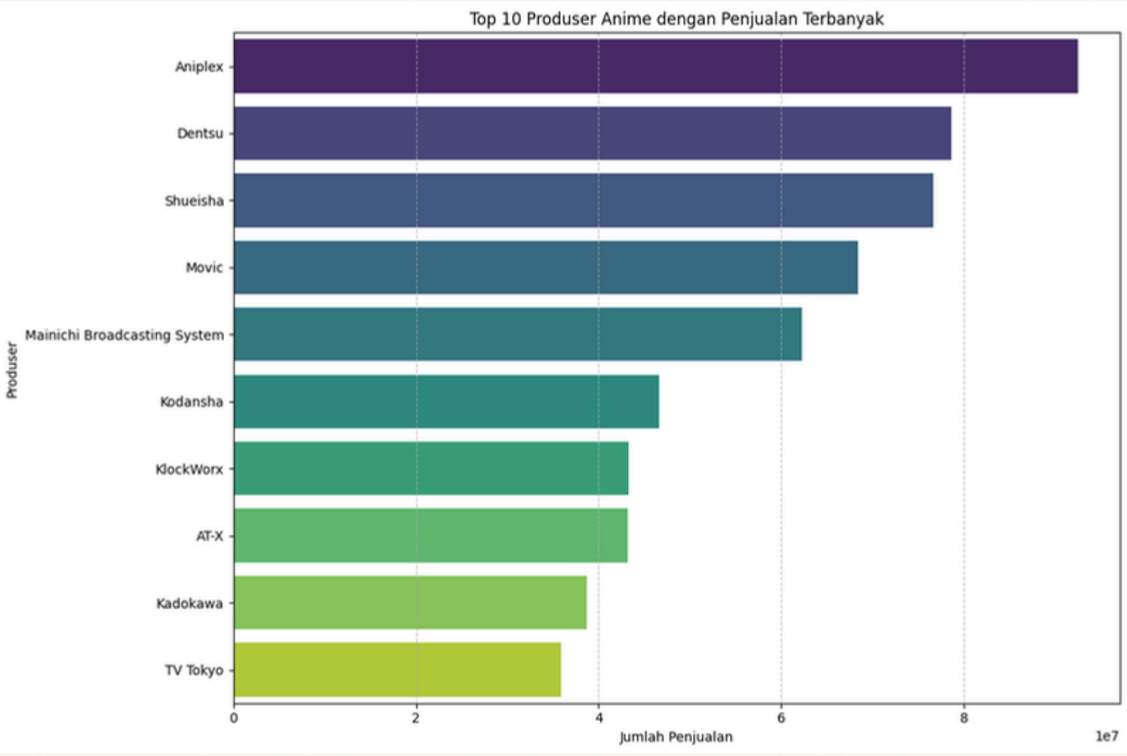
KASDAD - SILICON BALPEN

ANALISIS

1. Jumlah penayangan anime pada tahun awal j**umlahnya sangat sedikit**. Angka ini bisa jadi disebabkan oleh kelangkaan data historis, data yang tidak dimasukkan ke dalam dataset ini, atau jumlah anime yang tayang pada tahun itu memang berjumlah segitu.
2. Jumlah penayangan anime pada **tahun 2024 hanya berjumlah 7** yang sangat sedikit dibandingkan dengan **tahun 2023** dengan **jumlah 290** yang menjadi tahun p**uncak jumlah penayangan anime pada dataset ini.** Hal ini bisa jadi karena data penayangan anime pada tahun 2024 belum dimasukkan ke dataset, anime pada tahun 2024 belum semuanya tayang, atau dataset belum lengkap.
3. Jumlah anime yang tayang memiliki distribusi yang berbentuk **right skewed.**

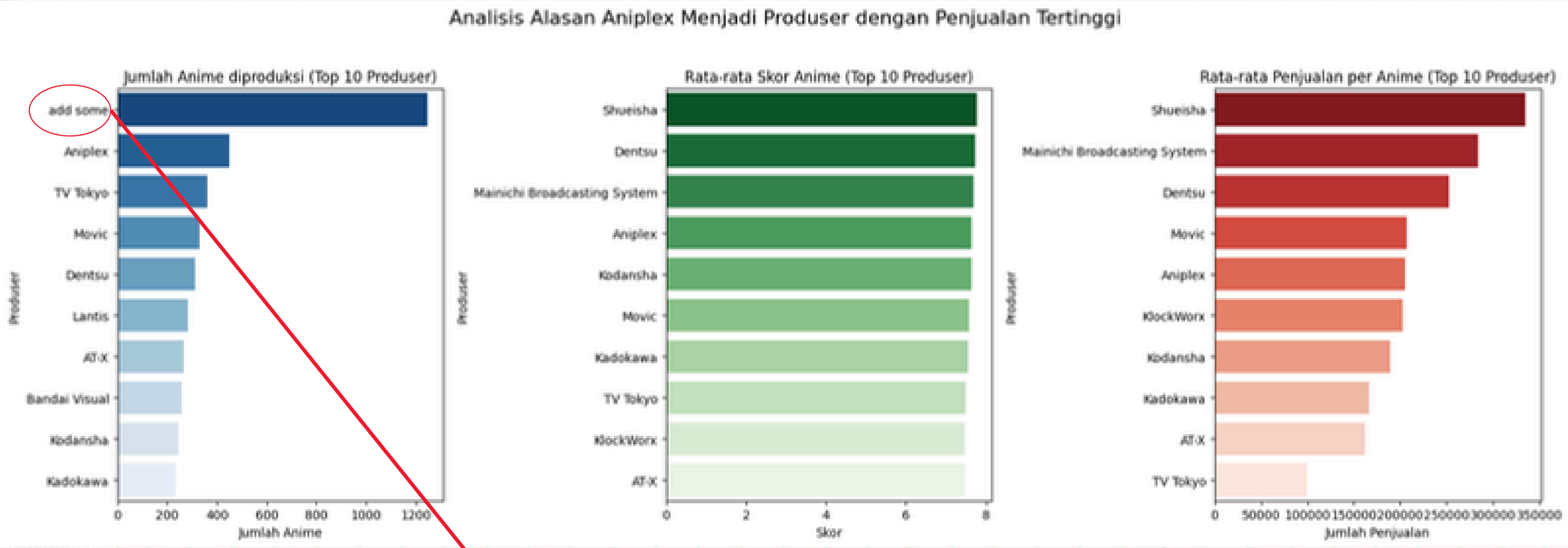
4. Produser apa yang menghasilkan penjualan anime terbanyak dari data tersebut? Mengapa produser tersebut yang menjadi penjual tertinggi?

	Produser	Jumlah Penjualan
0	Aniplex	92420614
1	Dentsu	78617305
2	Shueisha	76662179
3	Movic	68349123
4	Mainichi Broadcasting System	62233514
5	Kodansha	46607537
6	KlockWorx	43241580
7	AT-X	43218233
8	Kadokawa	38737294
9	TV Tokyo	35854062



Aniplex adalah produser dengan penjualan terbanyak

Mengapa Aniplex menjadi produser dengan penjualan terbanyak?



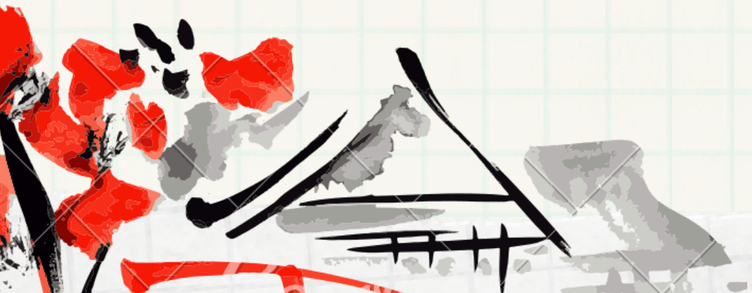
add some di sini berarti unknown

ANALISIS

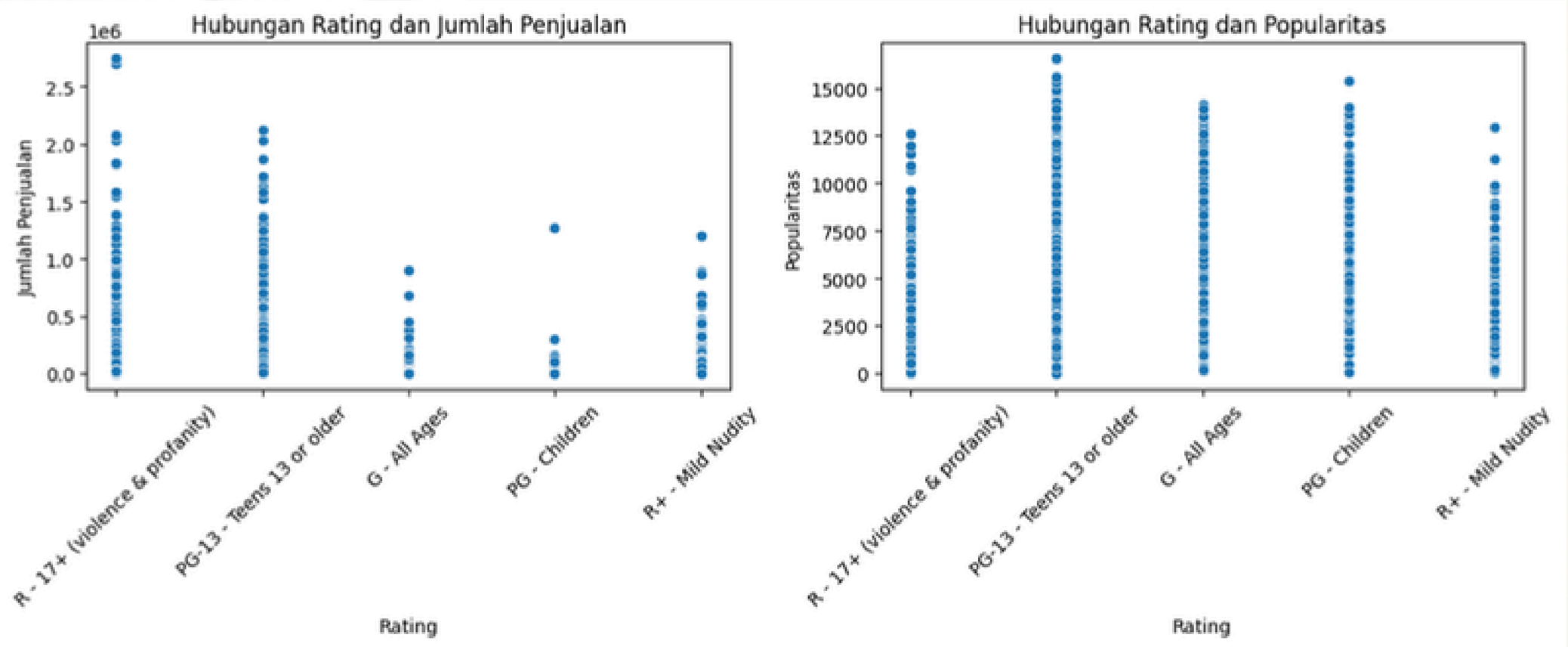
Alasan utamanya adalah kombinasi antara jumlah anime yang banyak dan penjualan rata-rata yang cukup tinggi. Meskipun Aniplex tidak menjadi yang terbaik di seluruh kategori, Aniplex unggul karena:

- Produser dengan anime terbanyak
- Konsistensi dalam penjualan
- Perpaduan Kualitas dan Kuantitas

Aniplex menjadi produser dengan penjualan total tertinggi karena Aniplex mampu **menggabungkan jumlah produksi anime yang besar dengan tingkat penjualan rata-rata yang stabil dan konsisten baik.**



5. Bagaimana hubungan antara Jumlah Penjualan dan Popularitas dengan Rating Anime?



KASDAD - SILICON BALPEN

ANALISIS

Koefisien korelasi antara Rating dan Jumlah Penjualan: 0.15
Koefisien korelasi antara Rating dan Popularitas: -0.33

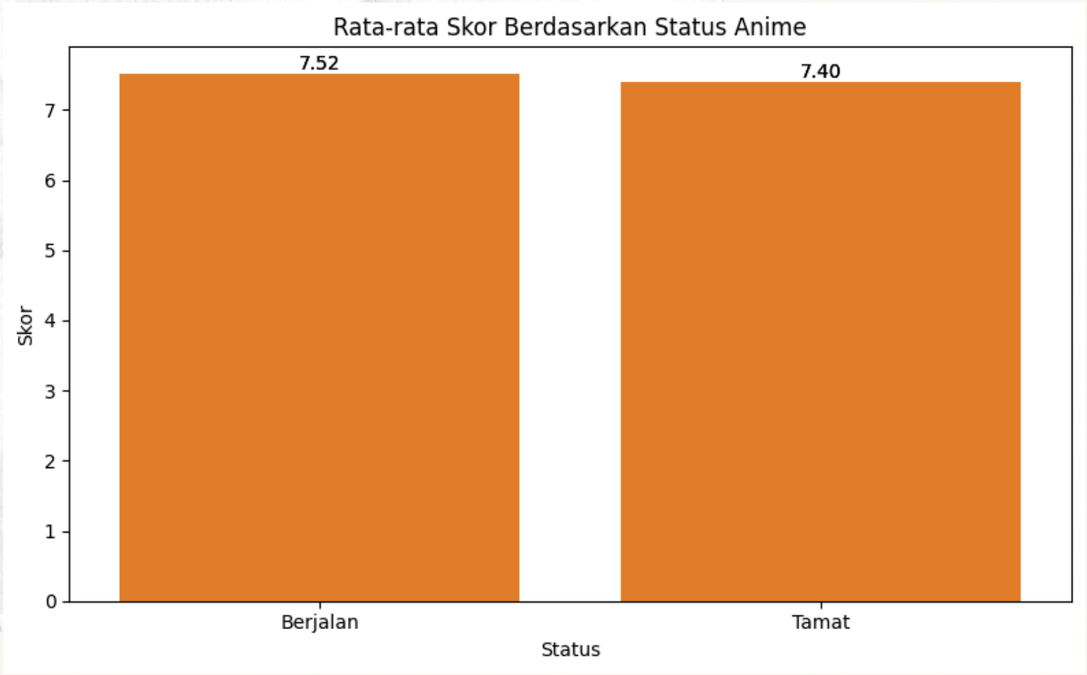
Rating vs Jumlah Penjualan

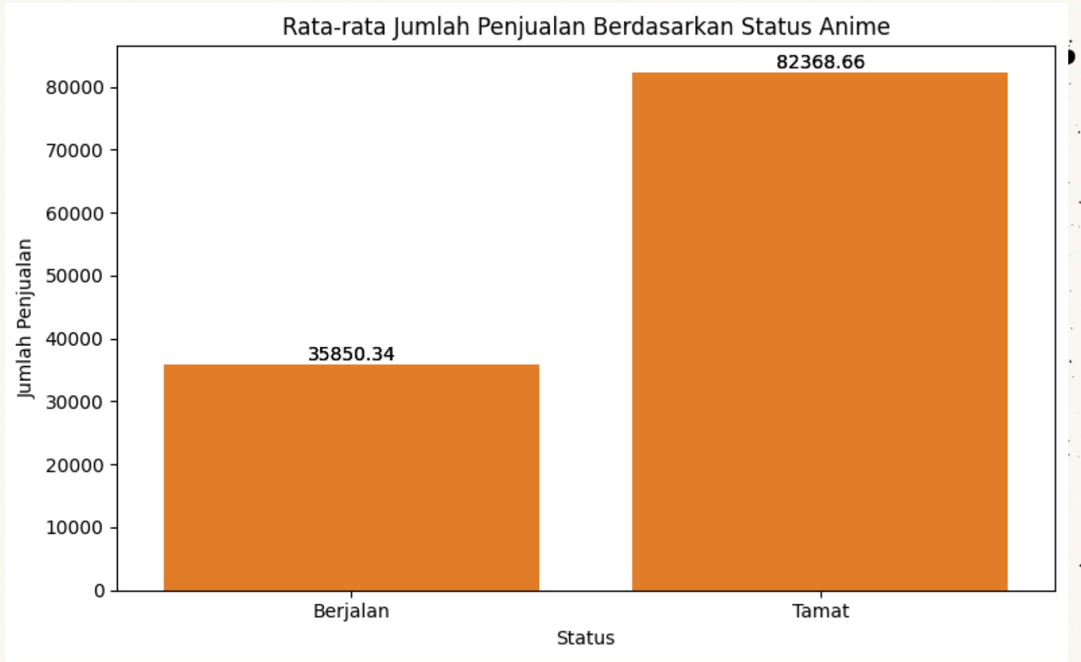
- Anime dengan rating PG-13 dan R – 17+ mendominasi penjualan tinggi, namun juga memiliki variasi yang luas.
- Rating anak-anak seperti PG – Children dan G – All Ages cenderung memiliki penjualan lebih rendah.
- Korelasi: +0.15 → **hubungan positif sangat lemah**, artinya rating dewasa sedikit berkorelasi dengan penjualan, namun bukan faktor utama.

Rating vs Popularitas

- PG-13 dan G – All Ages cenderung lebih populer dibanding rating dewasa.
- Anime dengan R – 17+ dan R+ memiliki popularitas yang lebih rendah dan bervariasi.
- Korelasi: -0.33 → **hubungan negatif sedang**, menunjukkan semakin dewasa rating, popularitas cenderung menurun.

6. Apakah anime dengan status berjalan dan status memiliki pengaruh terhadap skor dan penjualan anime tersebut?

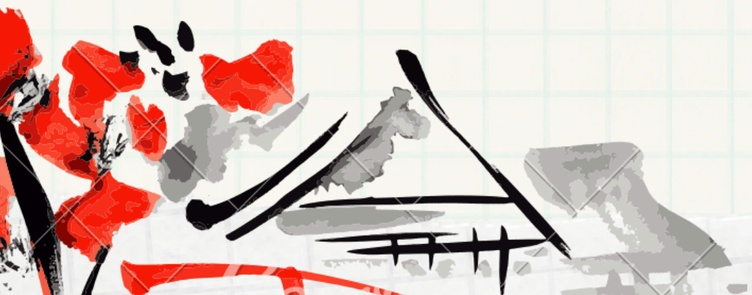




KASDAD - SILICON BALPEN

ANALISIS

- Perbandingan **Rata-rata skor** anime berdasarkan status tamat dan berjalan **tidak menunjukkan perbedaan besar**.
- hasil **rata-rata jumlah penjualan** dapat dilihat bahwa anime dengan status tamat memiliki rata-rata jumlah **penjualan yang lebih besar** dibandingkan dengan anime yang statusnya masih berjalan
- Hal ini bisa terjadi karena **preferensi penonton** yang lebih memilih untuk membeli atau menonton anime yang sudah tamat dibanding dengan anime yang masih berjalan **atau** hal ini bisa juga terjadi karena **anime yang tamat** sudah **berada di pasar lebih lama** dibandingkan dengan anime yang masih berjalan sehingga mereka sudah **mengumpulkan jumlah penjualan yang lebih banyak.**



DATA PREPROCESSING

Mengolah data agar bisa divisualisasikan dan diterima oleh model

- **Handle Missing Value**
- **Handle Duplication**
- **Handle Outliers**
- **Feature Engineering**
- **Standarization**
- **Imbalance Analysis**

HANDLE MISSING VALUES

```python
# Drop baris di data dengan Rating kosong
data = data.dropna(subset=["Rating"]).reset_index(drop=True)
```

Rating
Drop baris null

```python
# Isi missing value di kolom 'Premier' dengan kategori "Unknown"
data["Premier"] = data["Premier"].fillna("Unknown")
```

Premier
isi dengan “Unknown

```
Data Train:
          Total    Percent
Rating       20   0.333333
Premier    3090  51.500000
```

```
Data Train:
Tidak ditemukan missing value pada dataset
```

KASDAD - SILICON BALPEN

HANDLE DUPLICATE DATA

```
data_cleaned = data.drop_duplicates(subset=[col for col in data.columns if col != 'id'])
```

```
print(f"Jumlah data sebelum dibersihkan: {data.shape[0]}")
print(f"Jumlah data setelah dibersihkan: {data_cleaned.shape[0]}")
print(f"Duplikat yang dihapus: {data.shape[0] - data_cleaned.shape[0]}")
```

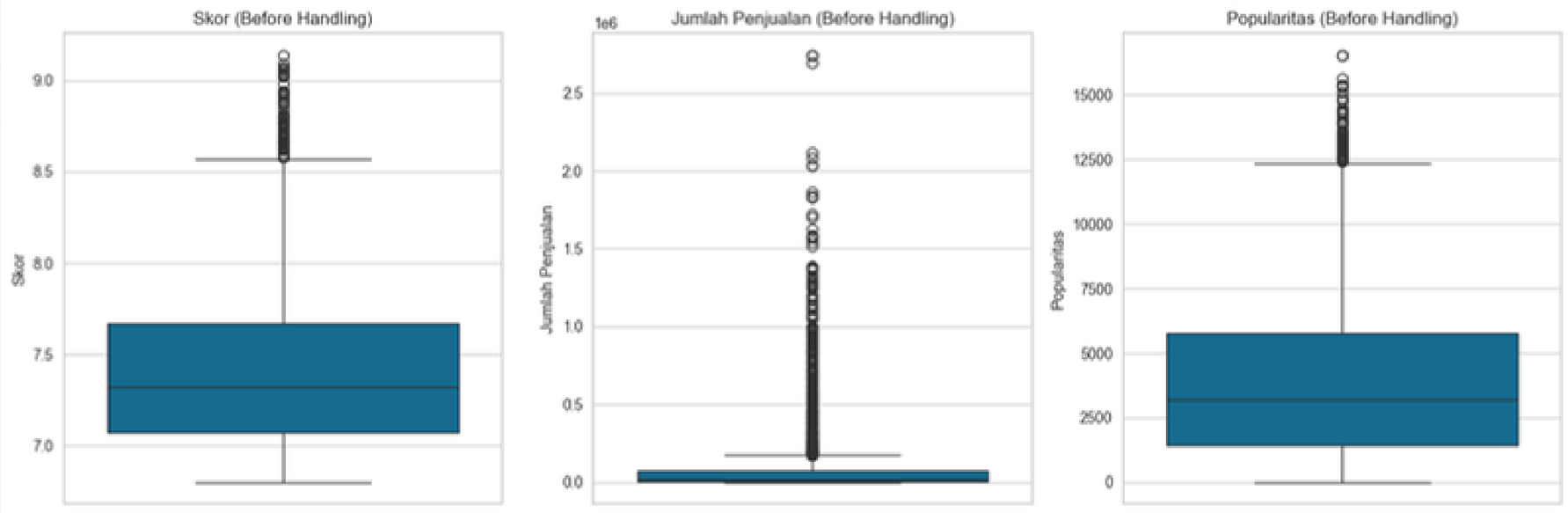
```
Jumlah data sebelum dibersihkan: 5980
Jumlah data setelah dibersihkan: 5412
Duplikat yang dihapus: 568
```

Ditemukan Data Duplikat yang memiliki ID Berbeda

HANDLE OUTLIERS

Log Transformasi
untuk mengurangi skewness

Winsorization (IQR)
untuk membatasi outlier ekstrem.



STANDARISASI

Fitur Durasi
Mengubah Durasi menjadi menit

Judul	Durasi
Ta Bu Dang Nuzhu Hen Duo NianSince I Wasn't th...	20 min. per ep.
Sayonara Watashi no Cramer Movie: First Touch	1 hr. 44 min.

Judul	Durasi (menit)
Ta Bu Dang Nuzhu Hen Duo NianSince I Wasn't th...	20.0
Sayonara Watashi no Cramer Movie: First Touch	104.0

Fitur Episode
Mengubah String menjadi int

Judul	Episode
Ta Bu Dang Nuzhu Hen Duo NianSince I Wasn't th...	16
Sayonara Watashi no Cramer Movie: First Touch	1

Judul	Episode
Ta Bu Dang Nuzhu Hen Duo NianSince I Wasn't th...	16.0
Sayonara Watashi no Cramer Movie: First Touch	1.0

STANDARISASI

Fitur Produser, Pemegang Lisensi, dan Studio
Mengubahnya menjadi String dan mengambil item pertamanya juga mengstandarisasi null tiap fitur

5	['Bandai Visual']	Bandai Entertainment	Bones
6	['add some']	None found, add some	Animation 21
7	['Heewon Entertainment']	None found, add some	None found, add some

5	Bandai Visual	Bandai Entertainment	Bones
6	Unknown	Unknown	Animation 21
7	Heewon Entertainment	Unknown	None found

FEATURE ENGINEERING

Fitur Tayang Mulai dan Tayang Selesai
Memecah fitur Waktu Penayangan untuk membantu memahami informasi fitur

Waktu Penayangan	Tayang Mulai	Tayang Selesai
Apr 22, 2023 to Oct 7, 2023	2023-04-22	2023-10-07
Jun 11, 2021	2021-06-11	2021-06-11
Nov 26, 2000 to May 29, 2001	2000-11-26	2001-05-29

Musim dan Tahun Premier
Didapatkan dari Fitur Tayang Mulai

Tayang Mulai	Musim Premier	Tahun Premier
2023-04-22	Semi	2023
2021-06-11	Panas	2021
2000-11-26	Gugur	2000

KASDAD - SILICON BALPEN

FEATURE SELECTION

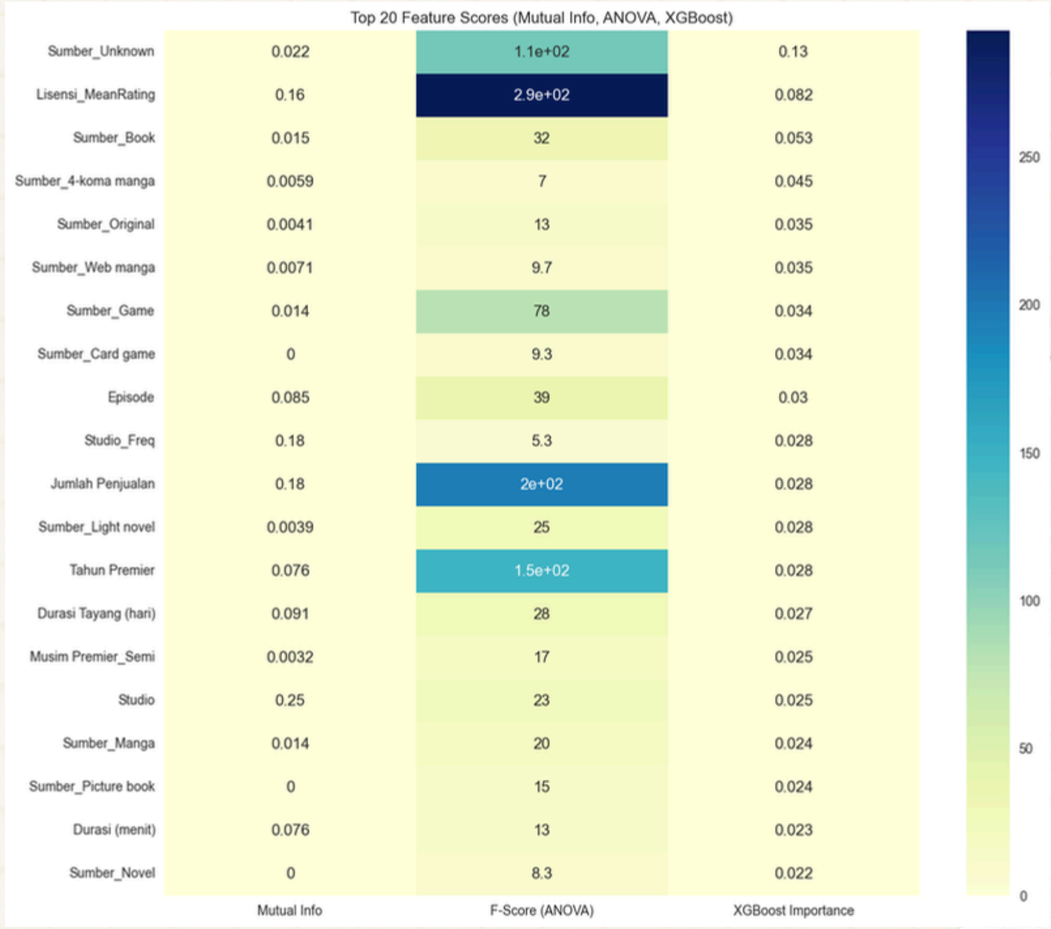
Feature Selection
Mengukur ketergantungan antara fitur dan target secara non-linear.

ANOVA F
Mengukur sejauh mana rata-rata nilai fitur berbeda antar kelas target.

Feature Importance XGBoost
Mengukur kontribusi masing-masing fitur berdasarkan performa model XGBoost.

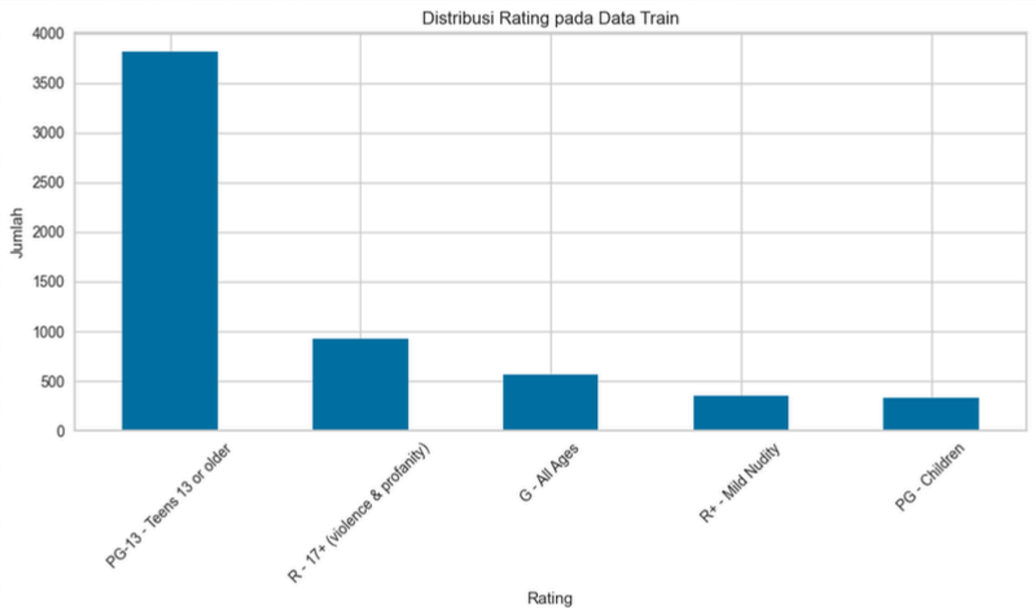
Fitur dipilih jika memenuhi salah satu dari kriteria berikut:

- Nilai Mutual Information > 0.05
- F-Score ANOVA > 10
- Importance dari XGBoost > 0.02



KASDAD - SILICON BALPEN

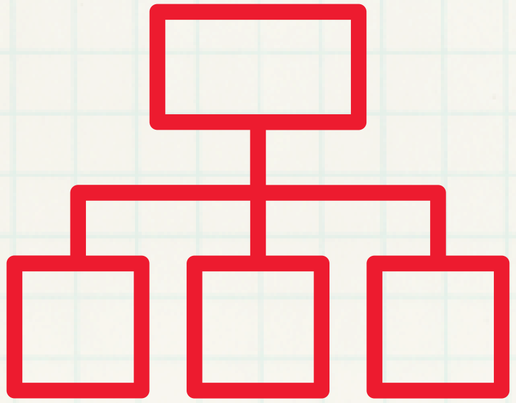
IMBALANCED ANALYSIS



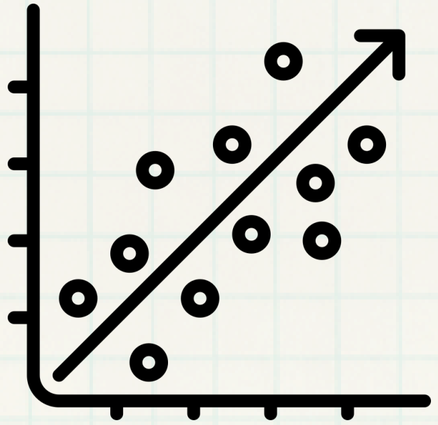
	Rating	Proportion	Proportion (%)
0	PG-13 - Teens 13 or older	0.635809	63.58
1	R - 17+ (violence & profanity)	0.153363	15.34
2	G - All Ages	0.092942	9.29
3	R+ - Mild Nudity	0.059497	5.95
4	PG - Children	0.058389	5.84

Data pada kelas PG-13 sangat mendominasi di dataset. Dominasi ini dapat menyebabkan model kesulitan untuk memprediksi kelas minoritas. Kami menangani multiclass imbalanced dengan **SMOTE** dan memanfaatkan **Stratified Kfold** untuk mempertahankan konsistensi dalam modeling.

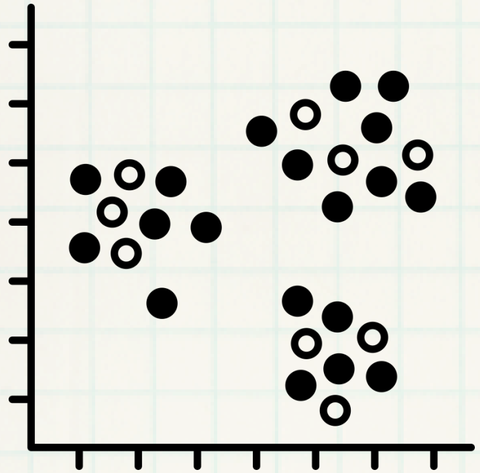
MODELLING



Classification



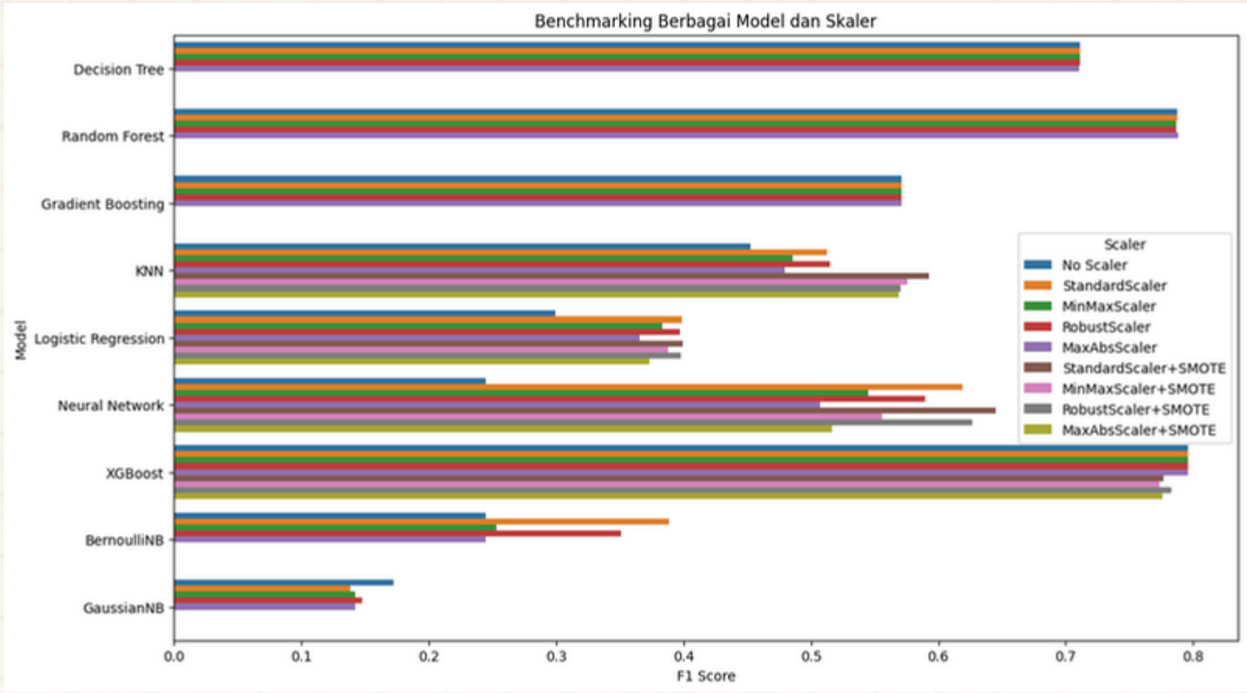
Regression



Clustering

KASDAD - SILICON BALPEN

BENCHMARK MODEL



Scaler	Model	F1 Score
No Scaler	XGBoost	0.795659
MaxAbsScaler	Random Forest	0.787883
StandardScaler	Decision Tree	0.711436
StandardScaler+SMOTE	Neural Network	0.644997
StandardScaler+SMOTE	KNN	0.592533
No Scaler	Gradient Boosting	0.571088
StandardScaler+SMOTE	Logistic Regression	0.399438
StandardScaler	BernoulliNB	0.388252
No Scaler	GaussianNB	0.171909

PERFORMANCE (TABEL SKOR SEMUA MODEL)

Random Forest	XGBoost	MLP	KNN	Decision Tree	Cat Boost
F-1 Score					
0.9993	0.8685	0.72930	0.72880	0.72471	0.9761262

KASDAD - SILICON BALPEN

RANDOM FOREST

f1 score base:
0.9993

hyperparameter:
n_estimators = 300
max_depth=20
min_samples_split=2
min_samples_leaf=1

f1 score tuning:
0.9975

Public score kaggle:
0.87539

KASDAD - SILICON BALPEN

MLP

f1 score base:
0.72815

hyperparameter:
**hidden_layer_sizes=(100, 50),
activation='relu',
solver='adam',
alpha=0.001,
learning_rate='constant',**

f1 score tuning:
0.7293

Public score kaggle:
0.82359

KASDAD - SILICON BALPEN

XGBOOST

f1 score base:
0.8037

hyperparameter:
**learning_rate=0.1,
max_depth=8,
n_estimators=200,
subsample=0.8,
colsample_bytree=1,**

f1 score tuning:
0.7956

Public score kaggle:
0.88322

KASDAD - SILICON BALPEN

CATBOOST

f1 score base:
0.95221

hyperparameter:
learning_rate = 0.1,
depth = 6,
stopping_rounds = 50,
class_weight = Balanced

f1 score tuning:
0.9761262

Public score kaggle:
0.8285383

CLASSIFICATION KAGGLE LEADERBOARD

#	Team	Members	Score	Entries	Last	Join
1	fourward		0.89040	39	17m	
2	**Silicon Balpen**		0.88983	67	13m	
	Your Best Entry! Your submission scored 0.82359, which is not an improvement of your previous score. Keep trying!					
3	Baghdad		0.88654	17	6h	
4	Datalicious		0.88653	76	2h	
5	BurhanBoost		0.86001	26	1h	
6	kasdaddy		0.85450	35	6h	
7	Kasdead		0.85158	36	15m	

KASDAD - SILICON BALPEN

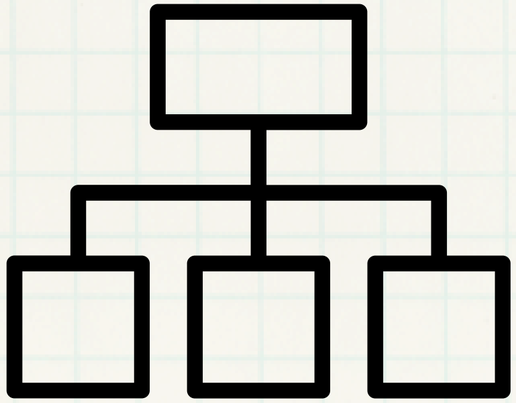
CLASSIFICATION KAGGLE LEADERBOARD

Public Private

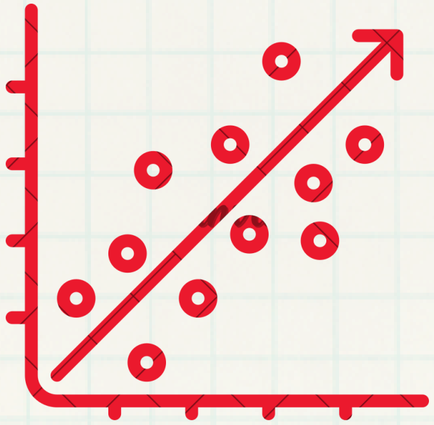
The private leaderboard is calculated with approximately 70% of the test data.
This competition has completed. This leaderboard reflects the final standings.

#	△	Team	Members	Score	Entries	Last	Solution
1	—	fourward		0.8827260	39	9d	
2	—	**Silicon Balpen**		0.8813409	67	9d	
3	—	Baghdad		0.8794553	19	9d	
4	—	Datalicious		0.8671337	77	9d	
5	▲ 1	kasdaddy		0.8575722	37	9d	
6	▼ 1	BurhanBoost		0.8559860	26	9d	
7	—	Kasdead		0.8502076	39	9d	
8	—	Data48		0.8352407	24	9d	
		benchmark_clasif.csv		0.6981062			

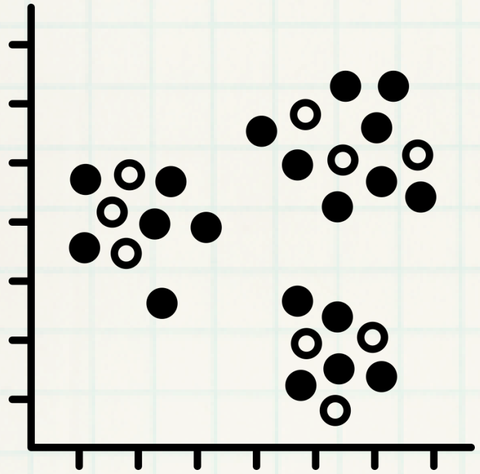
MODELLING



Classification

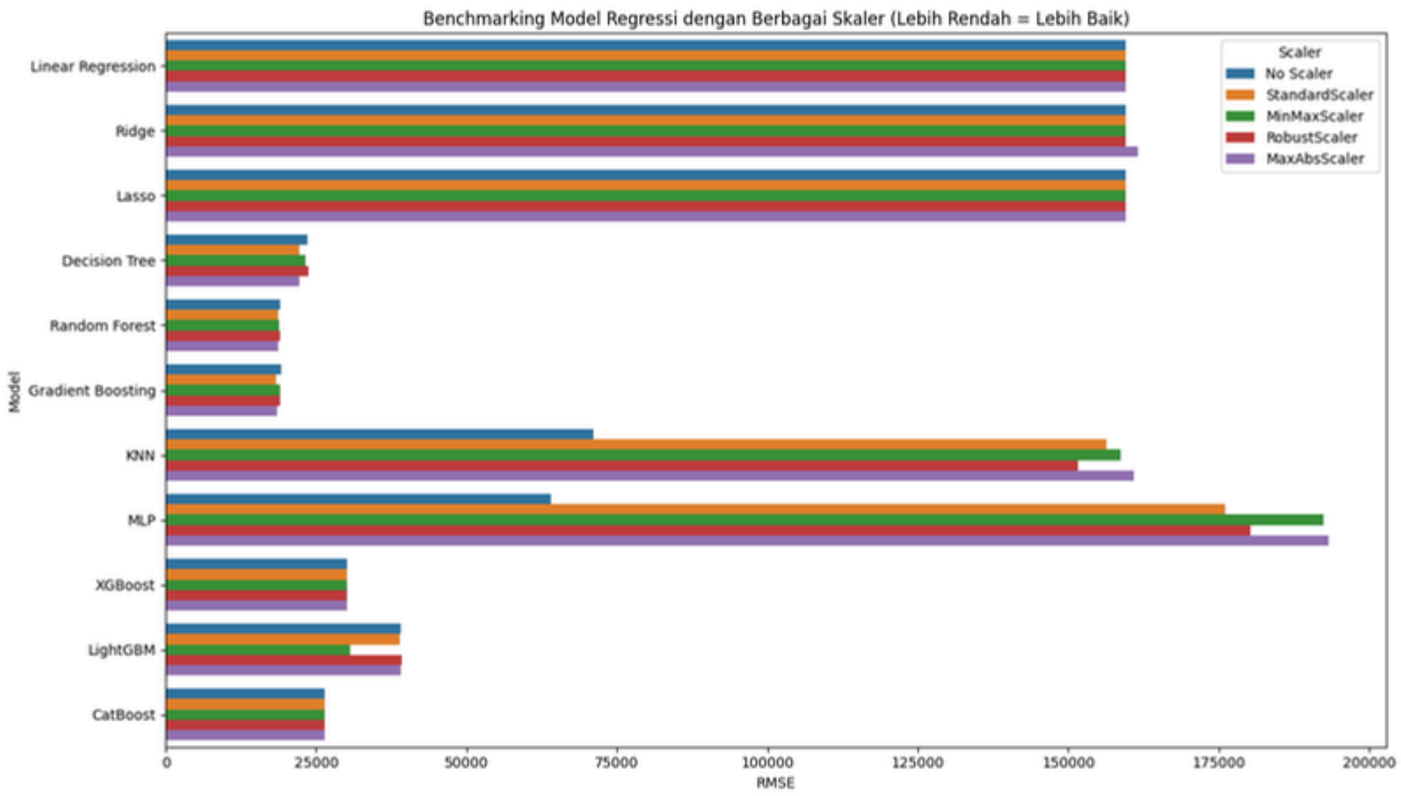


Regression



Clustering

BENCHMARK MODEL



Scaler	Model	RMSE
StandardScaler	Gradient Boosting	18379.937457
StandardScaler	Random Forest	18682.541799
MaxAbsScaler	Decision Tree	22165.856260
StandardScaler	CatBoost	26391.579861
No Scaler	XGBoost	30223.492188
MinMaxScaler	LightGBM	30720.277538
No Scaler	MLP	64039.456079
No Scaler	KNN	71126.376789
No Scaler	Ridge	159488.369613
MinMaxScaler	Lasso	159491.703345
No Scaler	Linear Regression	159492.015221

PERFORMANCE (TABEL SKOR SEMUA MODEL)

XGBoost	Lasso Regression	Ridge Regression	Random Forest	Cat Boost	Hist Gradient Boosting	lightGBM
R^2 Score						
0.9909	0.36117	0.39555	0.9866	0.9661	0.98489	**0.99391**

KASDAD - SILICON BALPEN

LIGHT GRADIENT BOOSTING MACHINE

Use parameter:
random_state=42,
n_estimators=500

R Squared:
1.000

Public score kaggle:
0.99391

KASDAD - SILICON BALPEN

RANDOM FOREST

Use parameter:
random_state=42, n_estimators=100

R Squared:
0.98701

Public score kaggle:
0.99338

KASDAD - SILICON BALPEN

RIDGE REGRESSION

Use parameter:
alpha=1

R Squared:
0.4499

Public score kaggle:
0.39555

KASDAD - SILICON BALPEN

XGBOOST

Use parameter:
**random_state=42,
n_estimators=2000,
max_depth=10,**

**learning_rate=0.03,
subsample=0.8,
colsample_bytree=0.8,
min_child_weight=5,
early_stopping_rounds= 10**

R Squared:
0.9909

Public score kaggle:
0.99274

KASDAD - SILICON BALPEN

CATBOOST

Use parameter:
iterations = 1000
learning_rate = 0.05
depth = 6
random_seed = 42

R Squared:
0.9661

Public score kaggle:
0.9936225

KASDAD - SILICON BALPEN

REGRESSION KAGGLE LEADERBOARD

#	Team	Members	Score	Entries	Last	Join
1	kasdaddy		0.9973569	25	2h	
2	BurhanBoost		0.9955064	19	1h	
3	Datalicious		0.9951917	43	6h	
4	fourward		0.9945564	42	2h	
5	**SiliconBalpen**		0.9939138	46	2h	
	Your Best Entry! Your submission scored 0.9913604, which is not an improvement of your previous score. Keep trying!					
6	Reza Taufiq Yahya		0.9938824	13	2h	
7	Baghdad		0.9938013	18	15m	
8	Kasdead		0.9935539	23	13m	

KASDAD - SILICON BALPEN

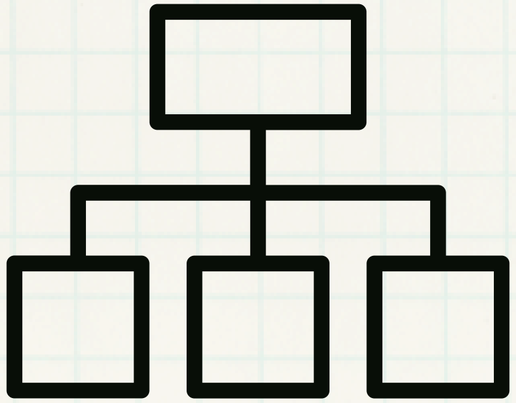
REGRESSION KAGGLE LEADERBOARD

Public Private

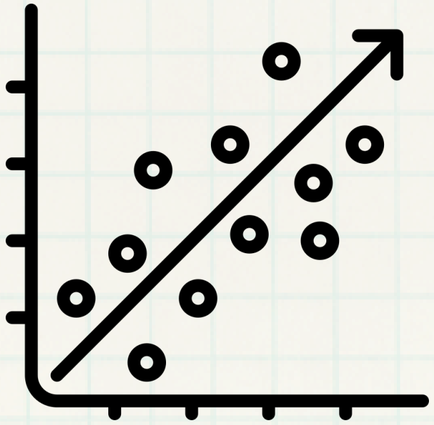
The private leaderboard is calculated with approximately 70% of the test data.
This competition has completed. This leaderboard reflects the final standings.

#	△	Team	Members	Score	Entries	Last	Solution
1	—	kasdaddy		0.9974039	25	9d	
2	▲ 5	Baghdad		0.9962515	18	9d	
3	▲ 2	**SiliconBalpen**		0.9961785	49	9d	
4	▲ 4	Kasdead		0.9961474	25	9d	
5	▼ 2	Datalicious		0.9959666	43	10d	
6	▼ 4	BurhanBoost		0.9956196	19	9d	
7	▼ 1	Reza Taufiq Yahya		0.9942458	13	9d	
8	▼ 4	fourward		0.9939803	42	9d	
		benchmark_regresi.csv		0.9887227			

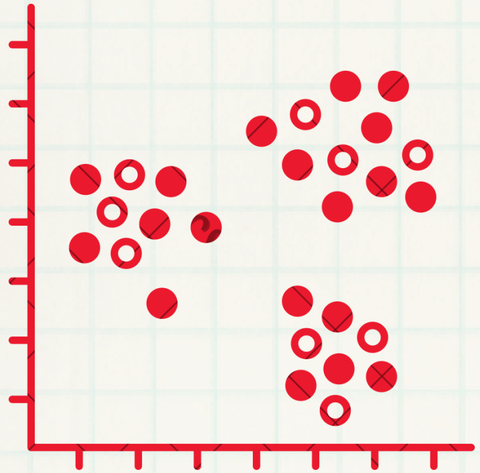
MODELLING



Classification



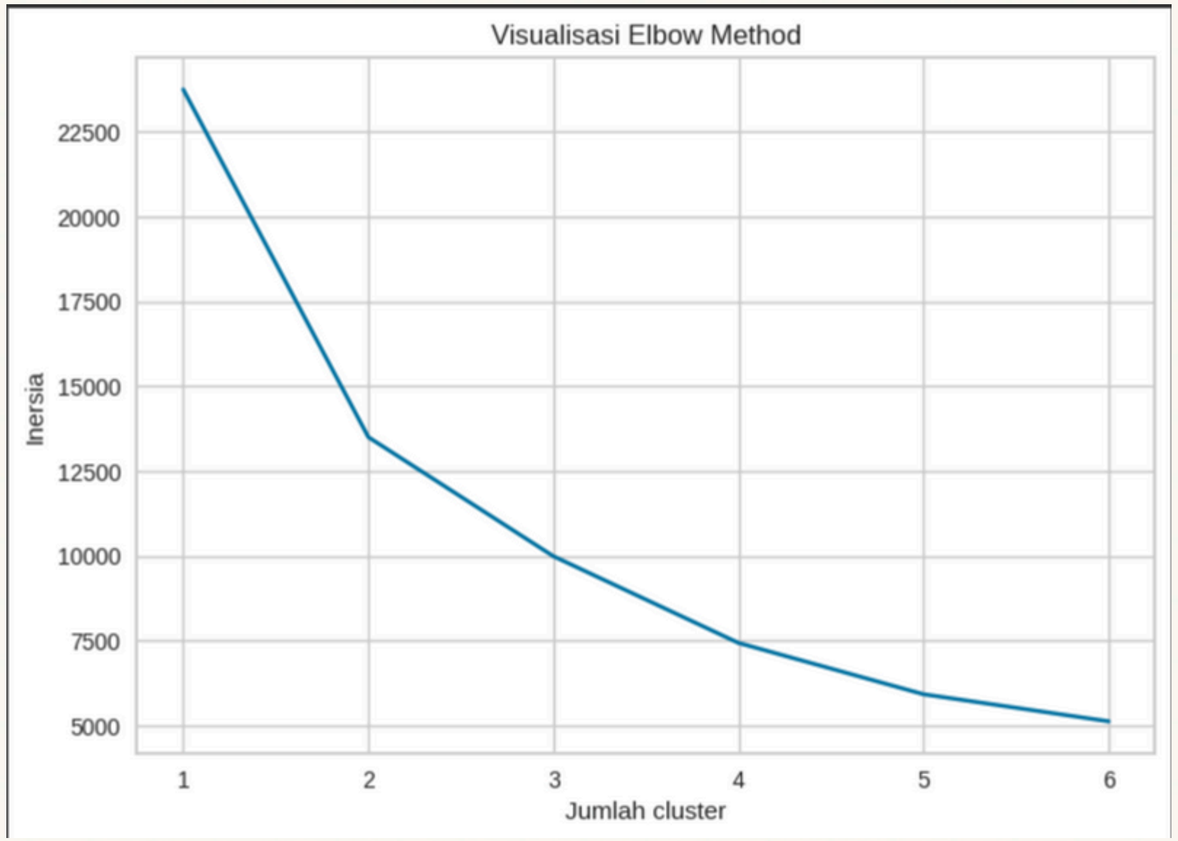
Regression



Clustering

KASDAD - SILICON BALPEN

KMEANS



Best n clusters: **2**

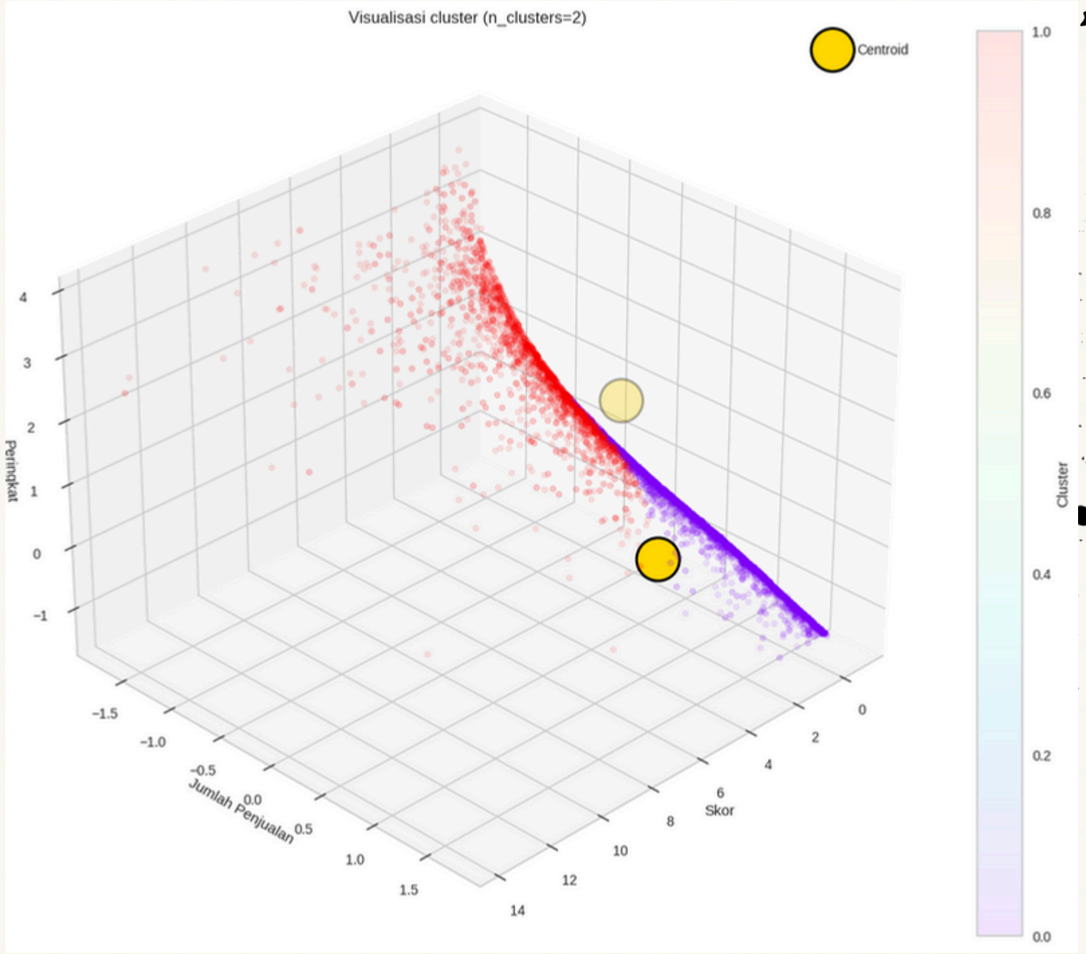
KASDAD - SILICON BALPEN

KMEANS

Best n clusters: **2**

Cluster pertama berisi anime-anime dengan skor user rendah, penjualan di bawah rata-rata, serta ranking yang baik

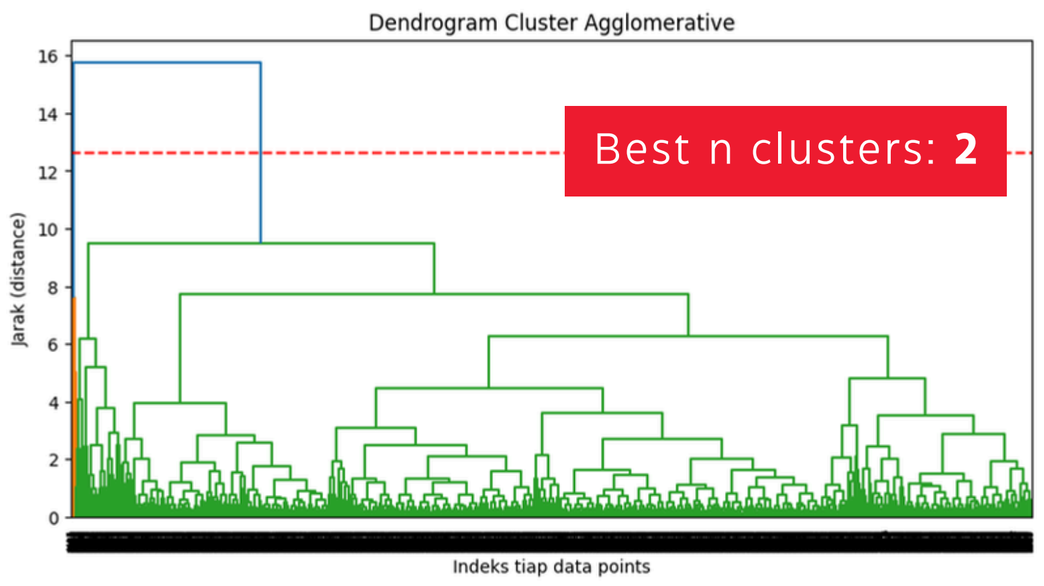
Cluster kedua diisi oleh anime-anime dengan skor user yang baik, penjualan yang tinggi, serta ranking yang rendah.



KASDAD - SILICON BALPEN

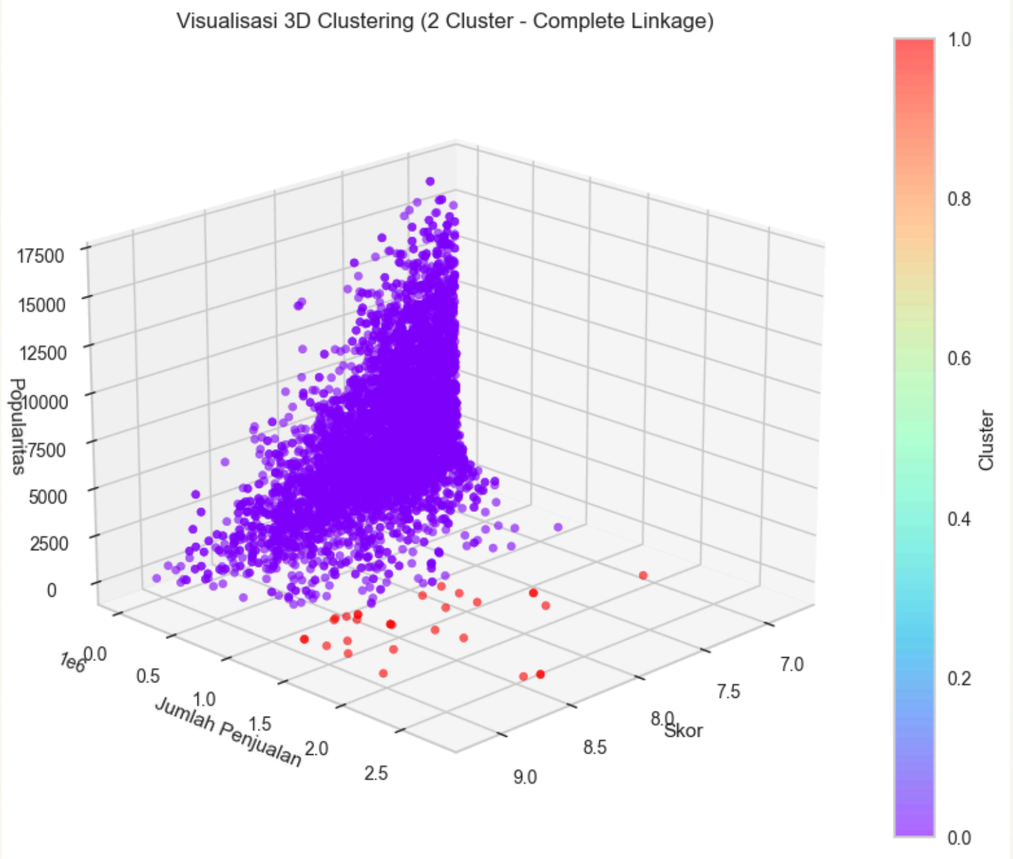
AGGLOMERATIVE CLUSTERING

Complete Linkage



Cluster besar untuk anime biasa

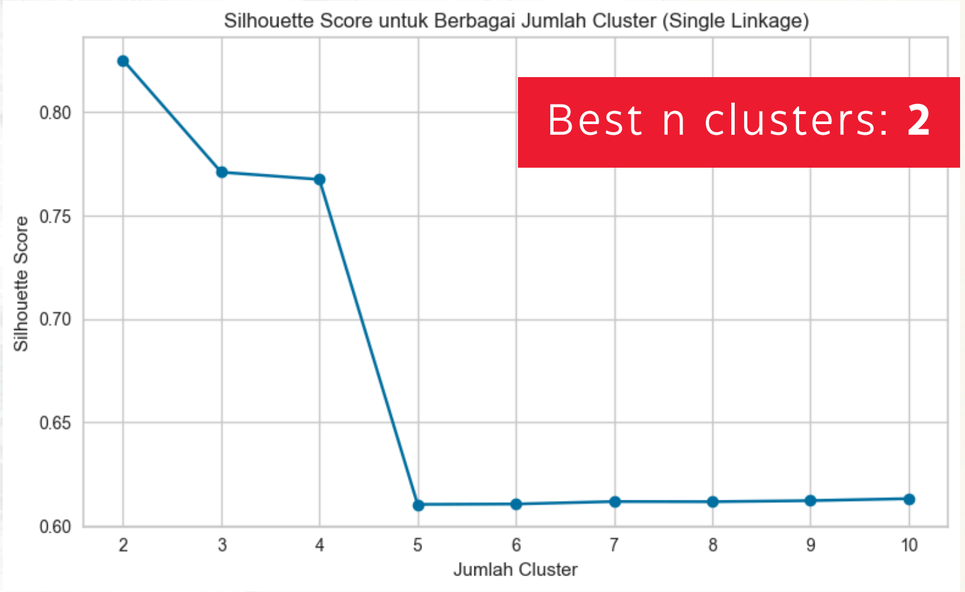
Cluster kecil khusus untuk hits dan legendaris



KASDAD - SILICON BALPEN

AGGLOMERATIVE CLUSTERING

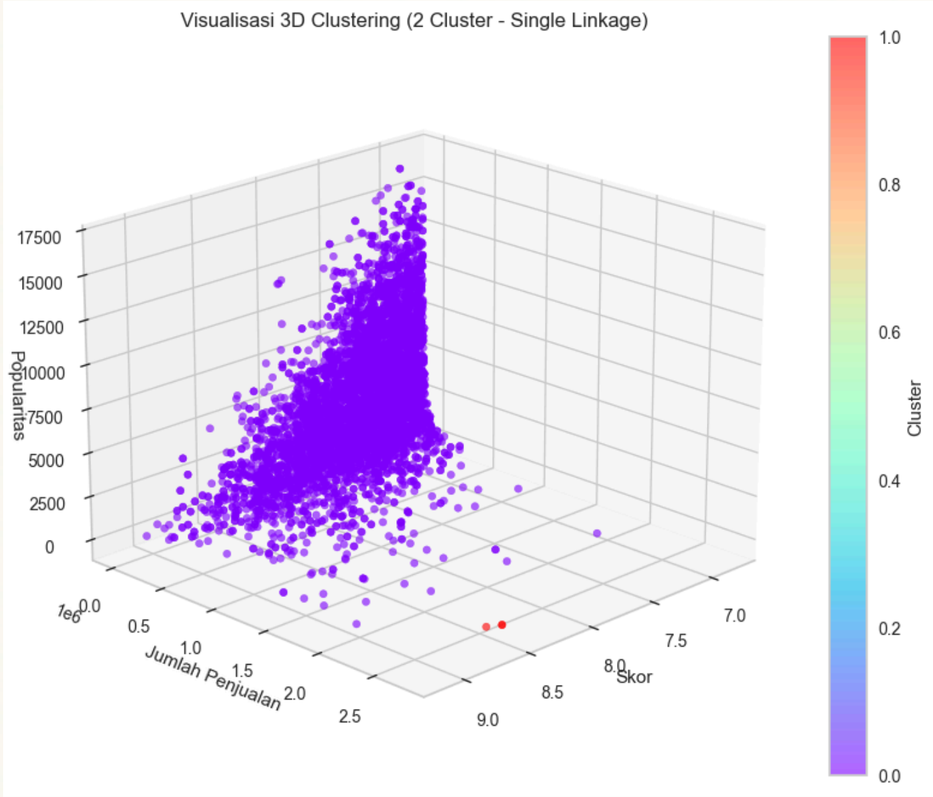
Single Linkage



Best n clusters: **2**

Cluster besar (umum) terdiri dari hampir semua anime pada dataset.

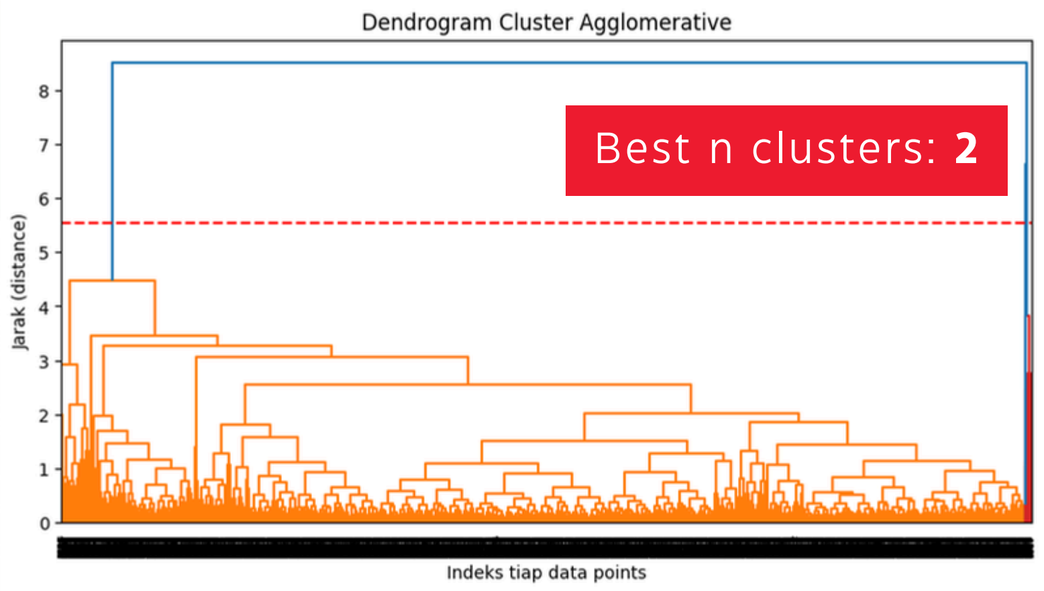
Cluster kecil berisi anime outlier



KASDAD - SILICON BALPEN

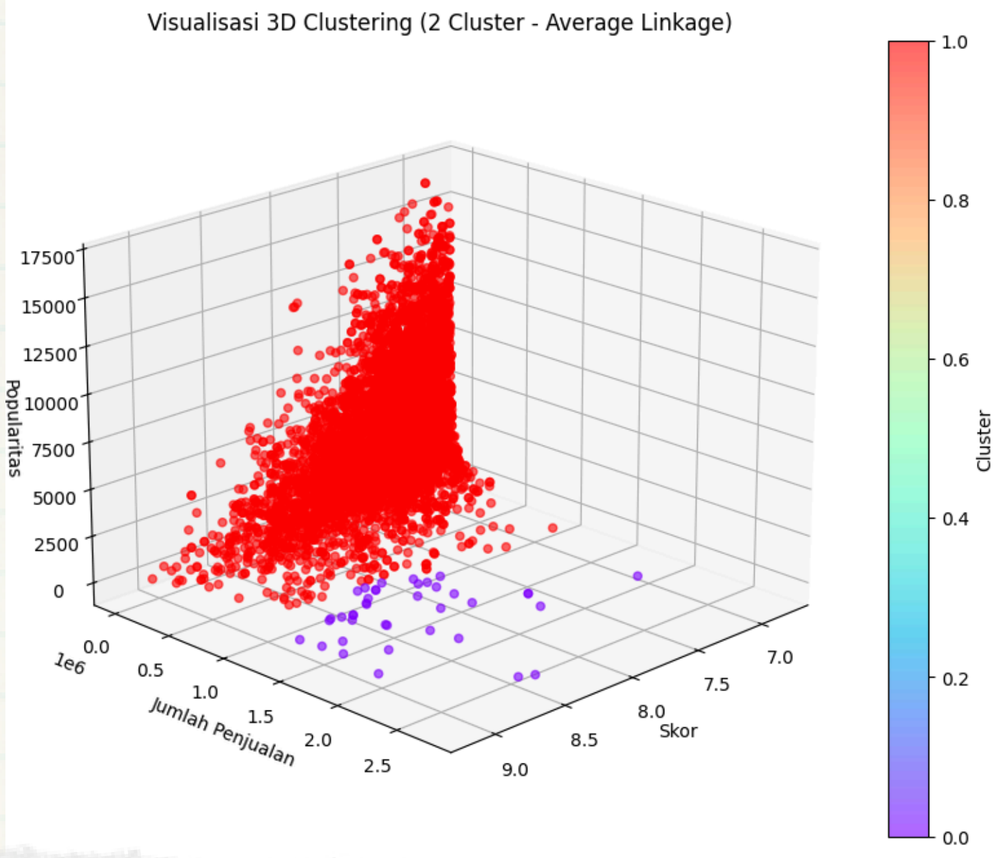
AGGLOMERATIVE CLUSTERING

Average Linkage



Cluster besar untuk anime biasa

Cluster kecil untuk anime hits dan legendaris



KESIMPULAN

EDA dilakukan untuk memahami karakteristik dataset melalui penggalian informasi terhadap dataset agar dapat membuat model yang baik.

Performa **Klasifikasi** terbaik dihasilkan oleh model **XGBoost** dengan nilai **F1 score 0.8037** dan public score **Kaggle** sebesar **0.88322**

Performa **Regresi** terbaik dihasilkan oleh model **LightGBM** dengan nilai **R squared** dan public score **Kaggle** sebesar **0.99391**

Clustering menunjukkan hasil yang konsisten meski metode berbeda. Terdapat **Dua cluster, satu segmen besar** terdiri dari anime dengan **performa dan popularitas rata-rata,** sementara **segmen kecil** terdiri dari **anime-anime luar biasa sukses** yang **mendominasi penjualan dan perhatian publik.**

SILICON BALPEN

THANK YOU!