

PREDICTION

Engagement Rate Short Video

Artificial Intelligence – Machine Learning

Team Member

Taufiq Rizan

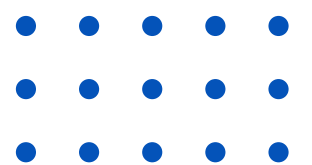
Adam Husein

Henry Kosasih

Devioletta Handayani

Ridho Ruhmanudin

Andre Pranata



BACKGROUND

Pada awal kepopuler media sosial (Facebook dan Twitter, instagram), jumlah pengikut (followers) adalah segalanya. Brand dan account saling bersaing untuk mendapatkan angka follower sebanyak mungkin. hal ini menimbulkan praktik "beli followers".

Kemudian platform media social mengubah urutan feed dari kronologis (berdasarkan waktu) menjadi algoritmik dengan **indicator utama engagement (interaksi)**

Jika sebuah content mendapat banyak interaksi dalam waktu cepat, algoritma akan menyebarkannya lebih luas sehingga metric follower sudah tidak relevan.



OBJECTIVE

Dunia digital sering kali menjadi hutan belantara yang penuh manipulasi sehingga dibutuhkan sebuah sistem yang mampu memberikan panduan dalam memahami data digital. Ketika semua pihak mulai dari pemilik brand, kreator, hingga platform memiliki acuan yang sama maka akan terbentuk ekosistem digital yang jauh lebih sehat, transparan, dan menguntungkan.



Brand & Advertiser

penghematan biaya dan waktu marketing.
ROI yang lebih efisien dan terukur



Content Creator

Pengakuan atas kualitas sebuah karya,
bukan sekadar popularitas atau kuantitas



Audiens & End User

Menerima paparan dan konsumsi konten
yang semakin relevan dan bermutu



Platform Media Social

menarik loyalitas audiens sekaligus
meningkatkan daya tawar ads spending

DATA & LIMITATION

✓ Short Video

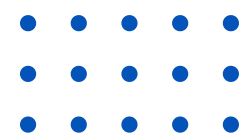
Data performance short-video dalam platform YouTube Shorts and TikTok pada tahun 2025

✓ Fitur Pre-Upload

Model focus menggunakan fitur pre-upload, yaitu informasi yang tersedia sebelum konten dipublikasikan

✓ Engagement Rate

Target yang diprediksi adalah engagement rate, yang didefinisikan sebagai rasio interaksi terhadap View





DATA INFORMATION

✓ Data Resource

Data performance short-video dalam platform YouTube Shorts and TikTok pada tahun 2025

✓ Target Definition

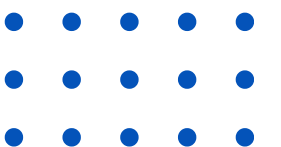
$\text{Engagement_total} = \text{likes} + \text{comments} + \text{shares} + \text{saves}$

$\text{Engagement_rate} = \text{engagement_total} / \text{views}$

✓ Data Frame

48,079 entries dan 58 column

data types : float64 (15), int64 (14), object (29)



PROCESS DEVELOPMENT



Exploratory Data Analysis

Memahami struktur, karakter,
distribusi dan data insight



Data Preprocessing

mengubah data mentah
menjadi data untuk modelling



Model Development

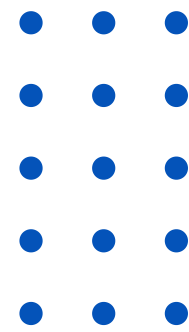
Membangun, Melatih, dan
mengevaluasi model



Model Deployment

mengubah model ML menjadi
solusi nyata dan useable

EDA

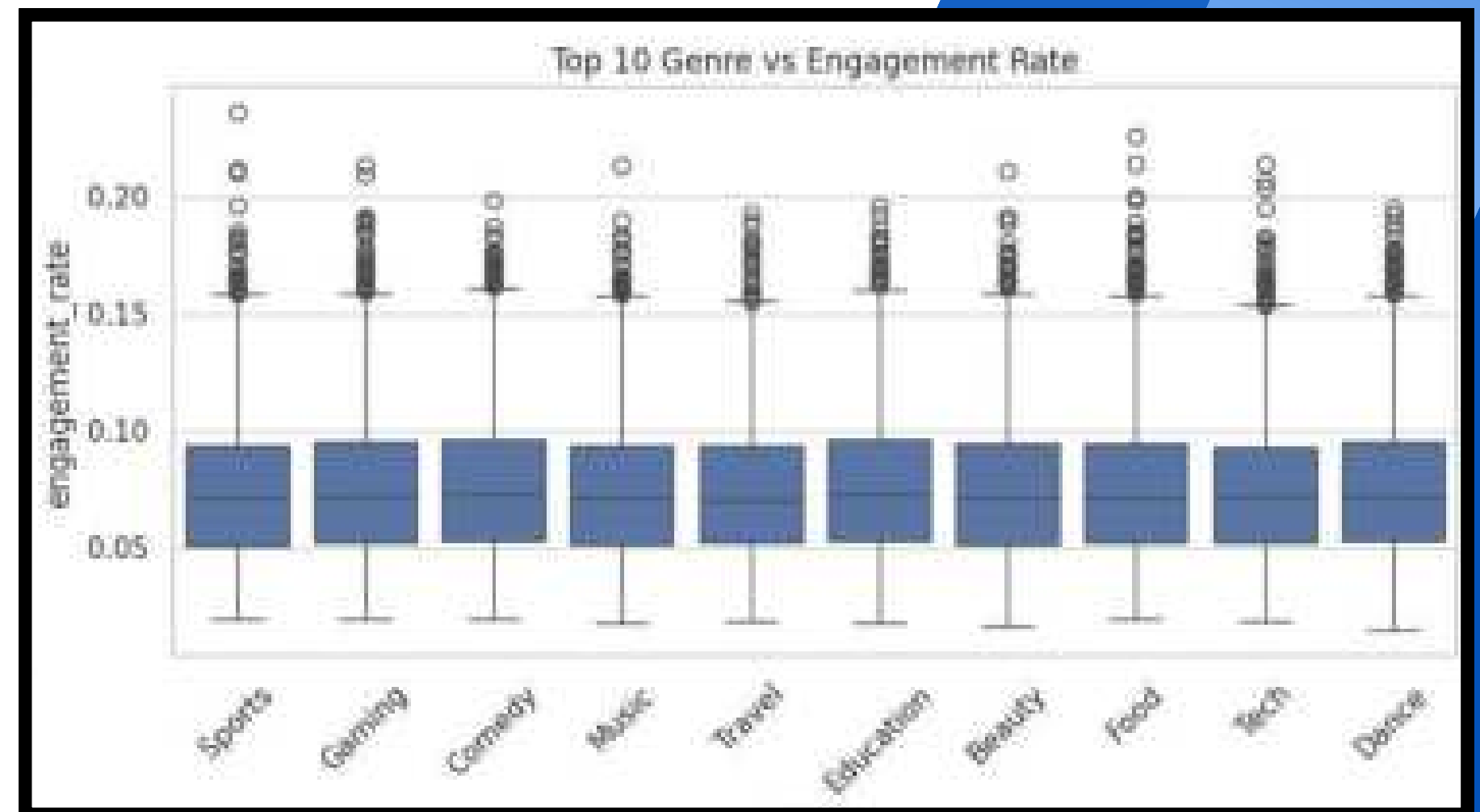
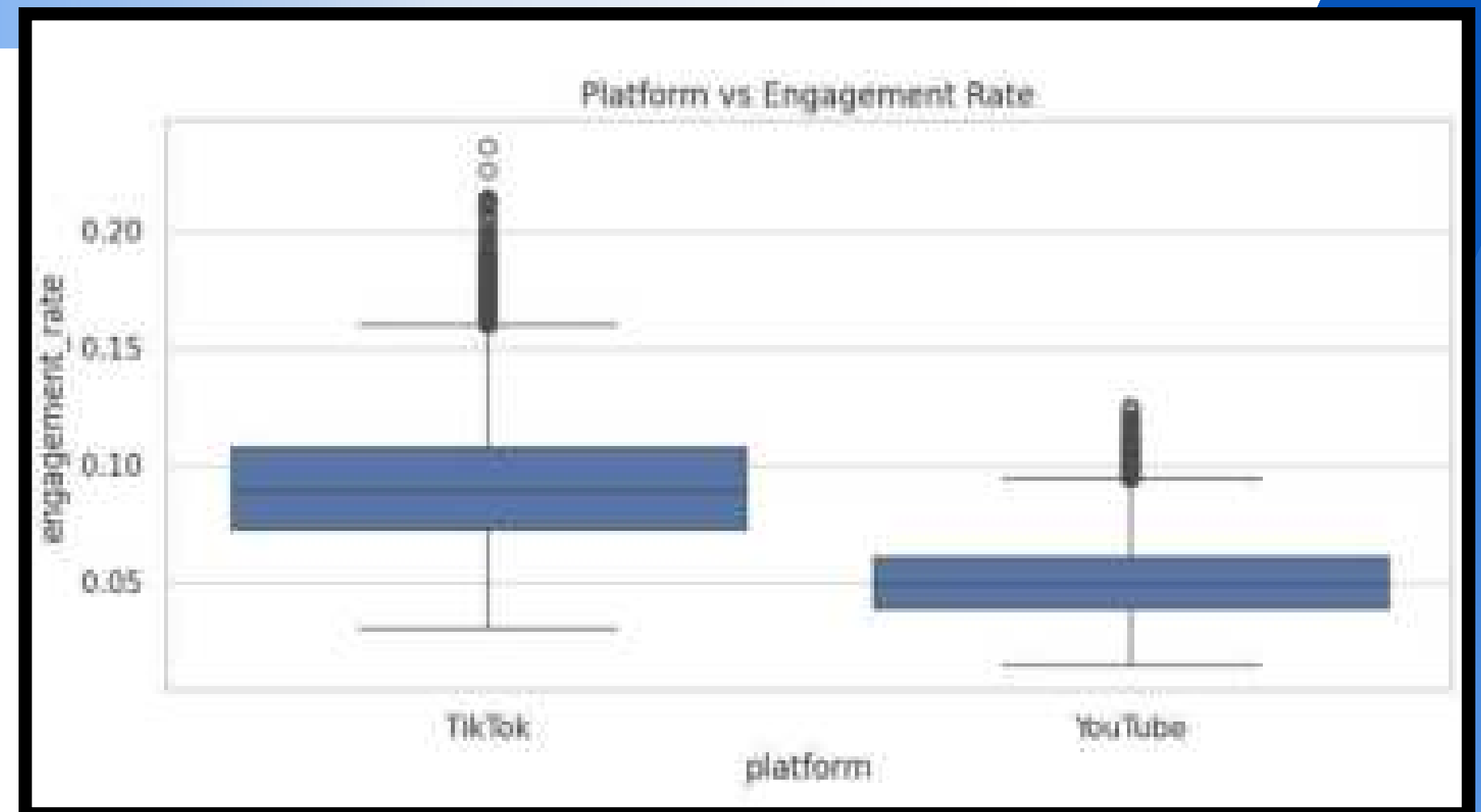


✓ Platform Type

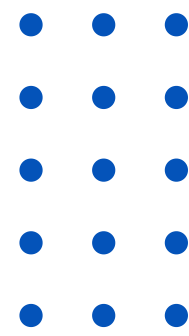
- TikTok memiliki engagement rate median jauh lebih tinggi dibanding YouTube
- TikTok unggul di engagement cepat
- YouTube lebih cocok long-term value

✓ Top 10 Genre

- Median engagement antar genre relatif mirip, Tidak ada genre yang ekstrem unggul atau tertinggal
- outlier muncul hampir semua genre
- Genre bukan faktor penentu utama



EDA

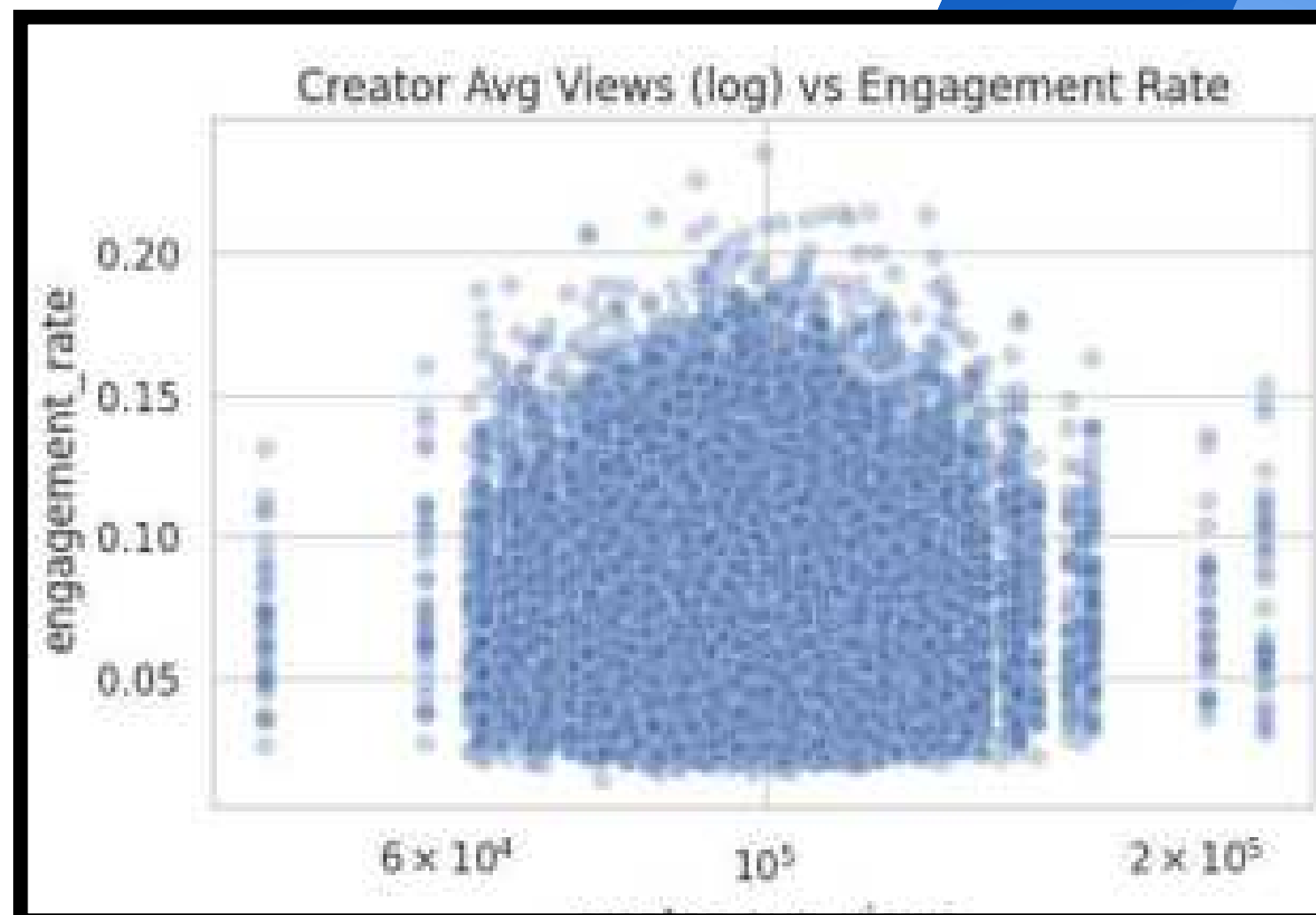
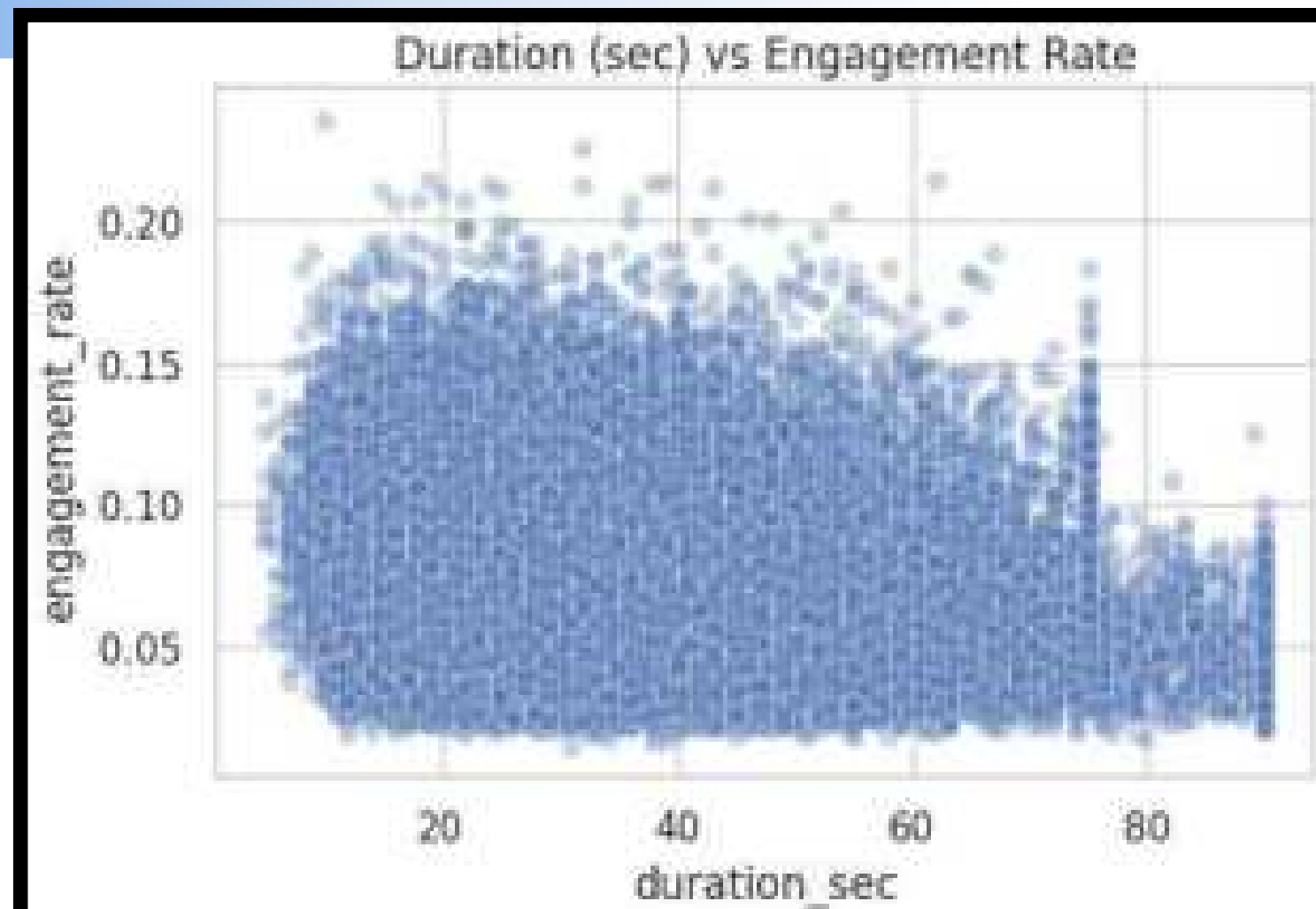


✓ Duration (sec)

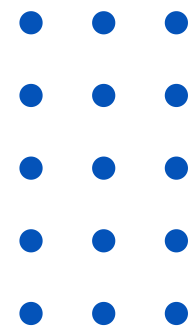
- Engagement rate tertinggi terkonsentrasi pada durasi pendek–menengah ($\pm 15\text{--}40$ detik)
- Semakin panjang durasi (>60 detik), engagement cenderung menurun

✓ Creator AVG View

- Engagement tinggi banyak muncul pada creator dengan average views menengah
- Creator dengan views sangat tinggi engagement lebih stabil



EDA

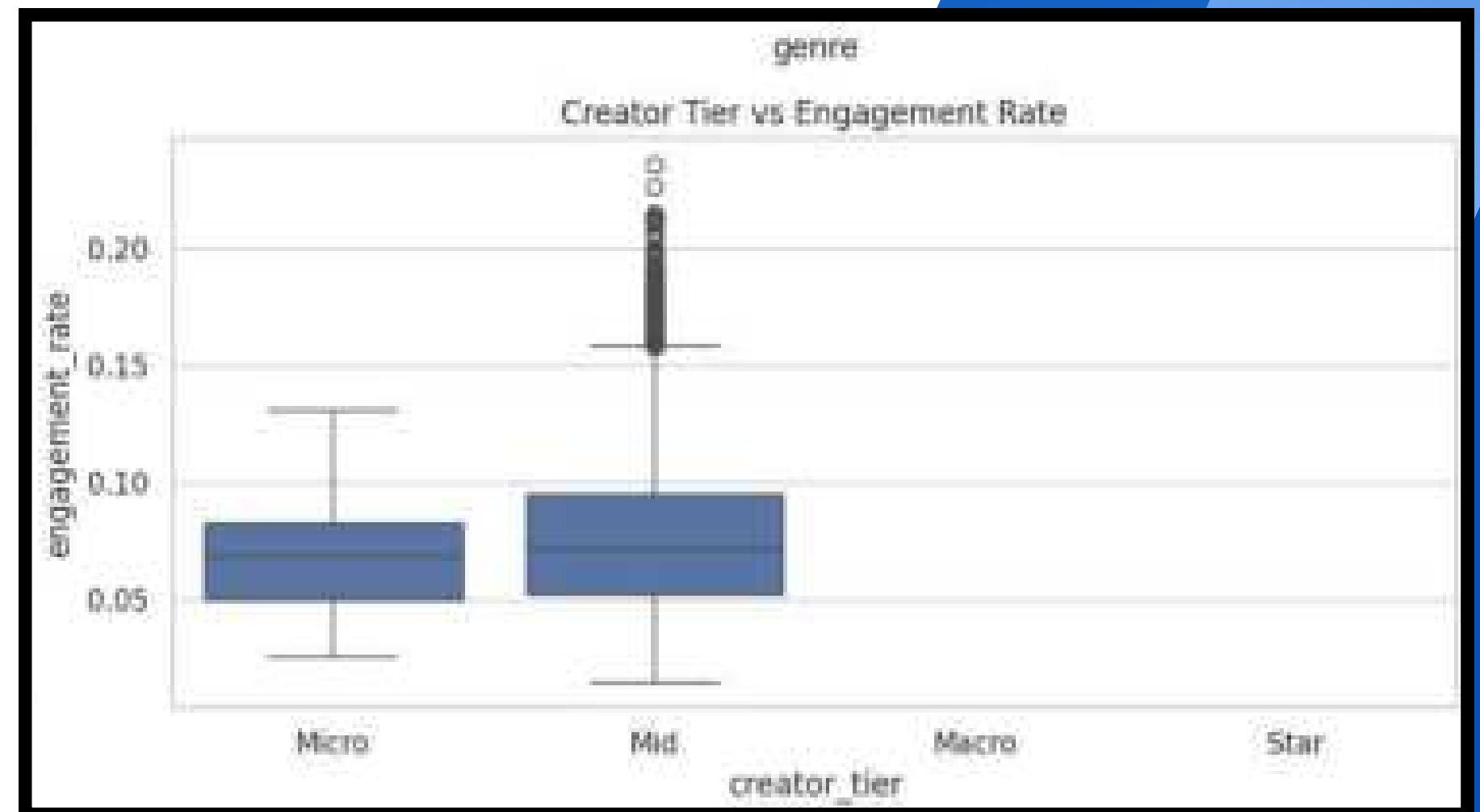
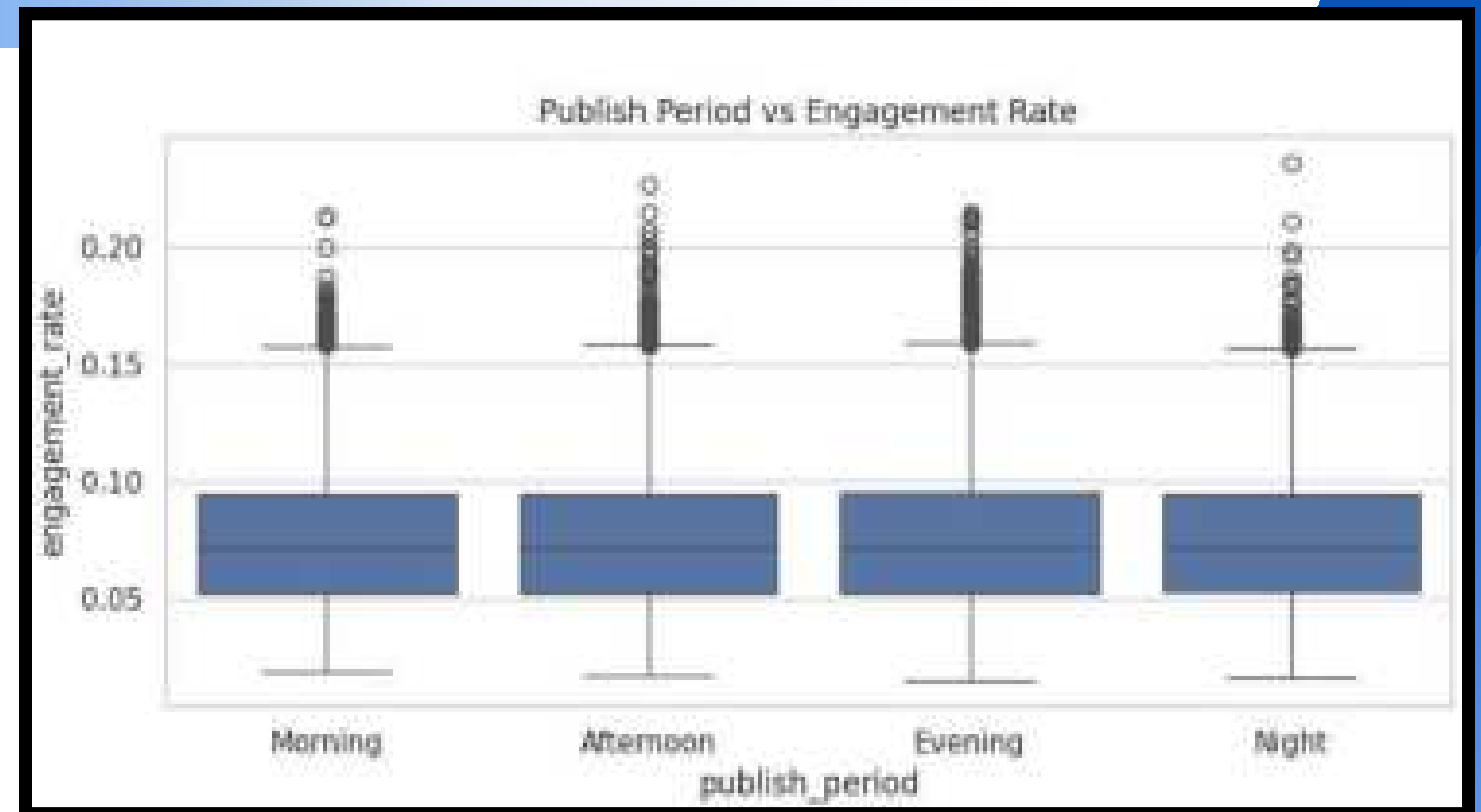


✓ Publish Period

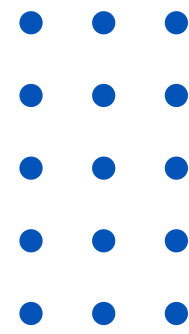
- Morning, Afternoon, Evening, Night → median hampir sama
- Tidak ada waktu posting yang dominan signifikan
- Algoritma lebih content-driven daripada time-driven

✓ Creator Tier

- Mid creator punya median engagement tertinggi
- Macro creator lebih stabil tapi kurang eksplosif

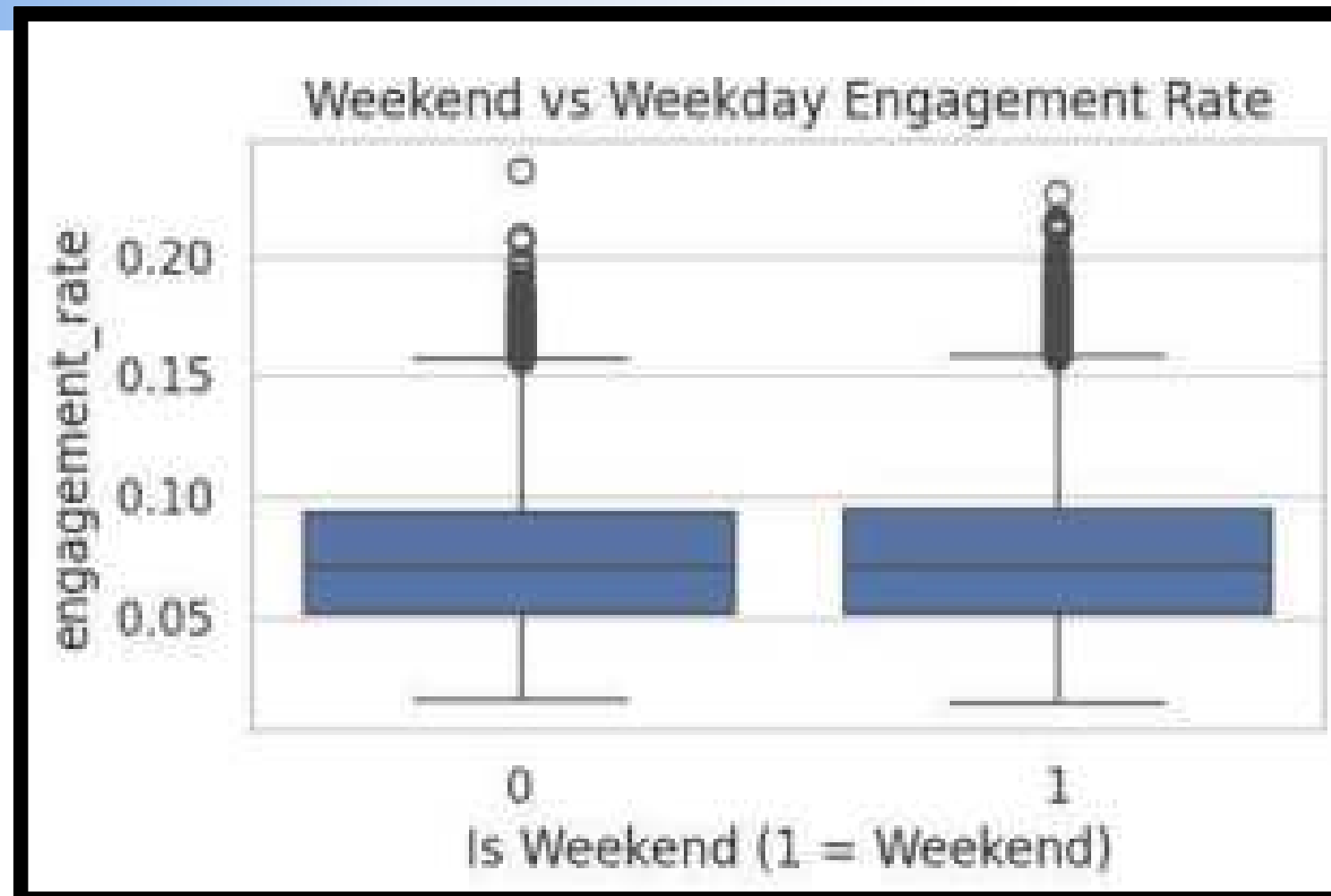


EDA



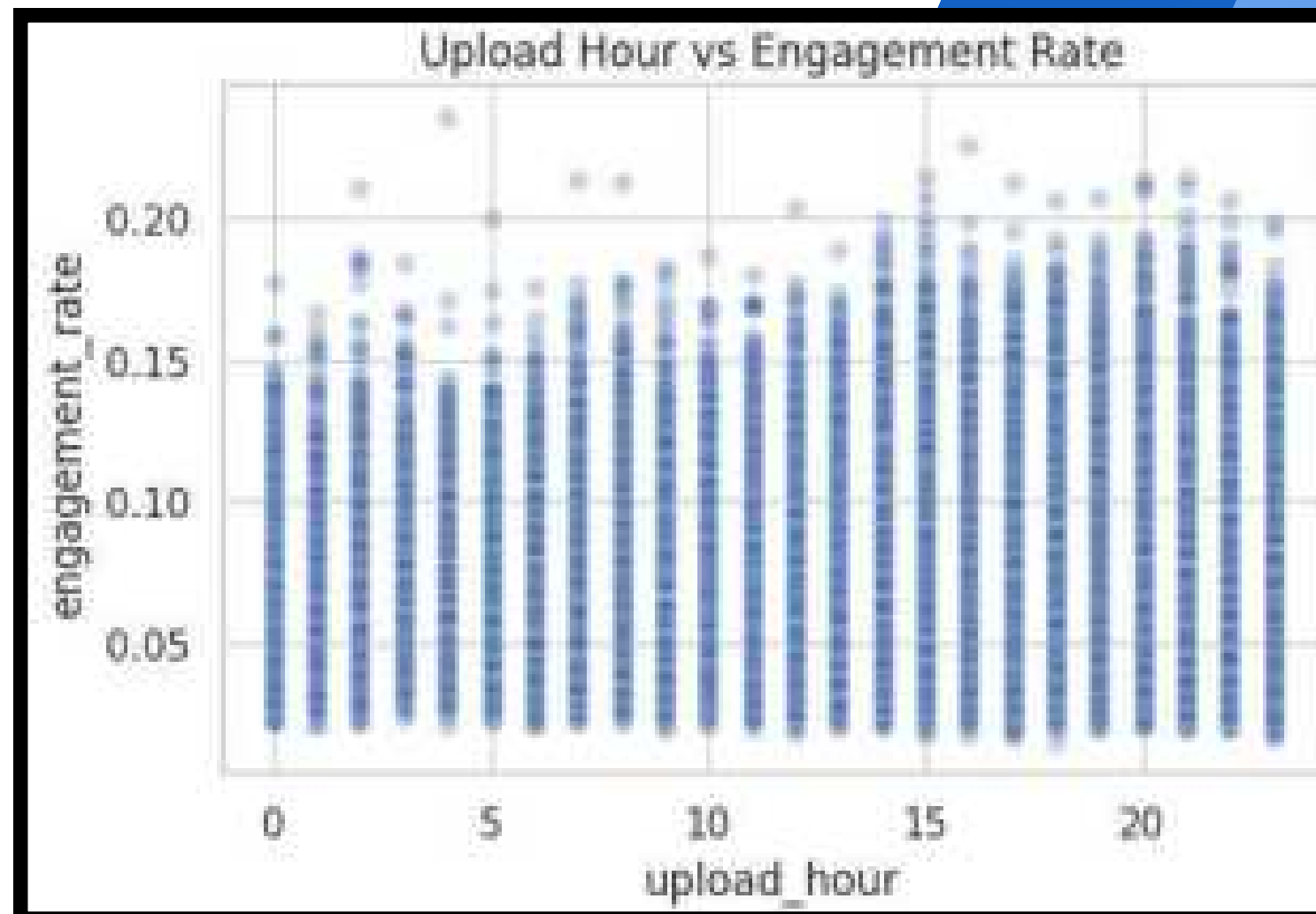
✓ Weekend – Weekday

- Weekend memiliki median engagement sedikit lebih tinggi
Namun perbedaannya tidak drastis

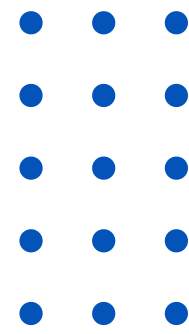


✓ Upload Hour

- Engagement tersebar merata di hampir semua jam
- Sedikit kecenderungan engagement lebih tinggi pada (\pm 16:00–22:00)



EDA

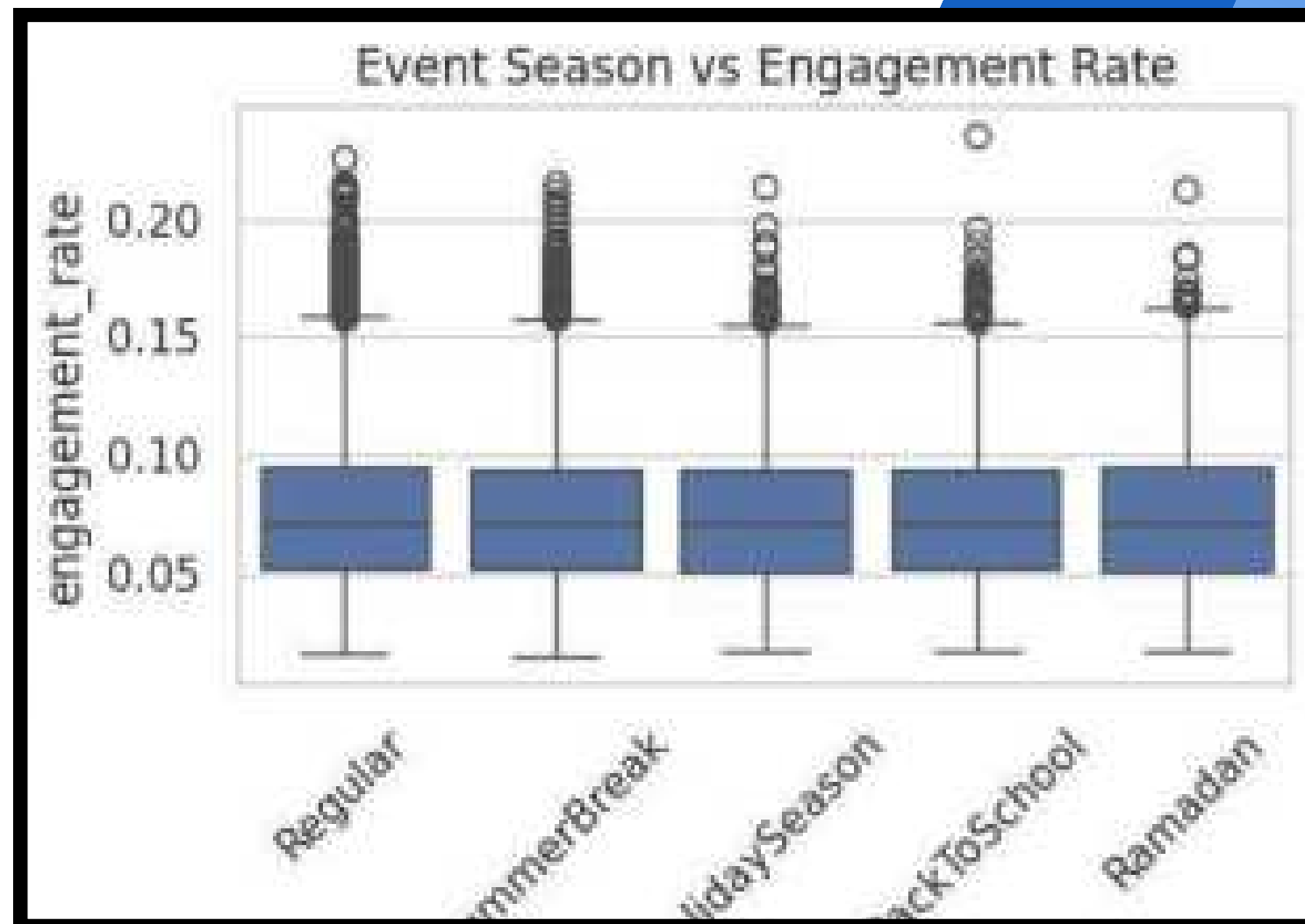
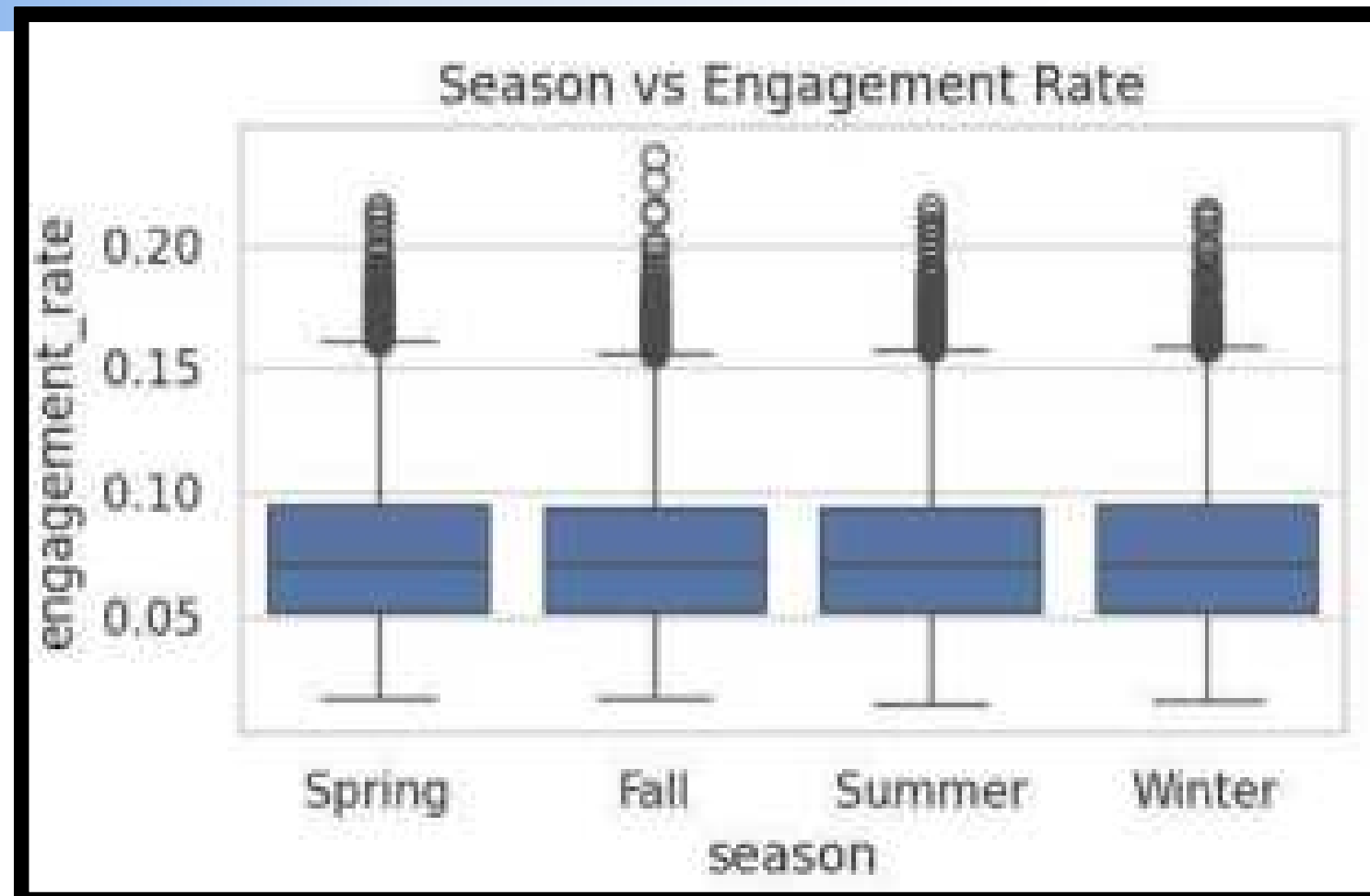


✓ Season

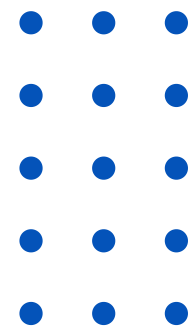
- Median engagement rate di semua musim relatif mirip. Tidak ada musim yang konsisten jauh lebih unggul
- Outlier tinggi muncul di semua musim

✓ Even Season

- Ramadan cenderung memiliki median engagement sedikit lebih tinggi disusul Holiday season
- Varian engagement masih cukup besar di semua event

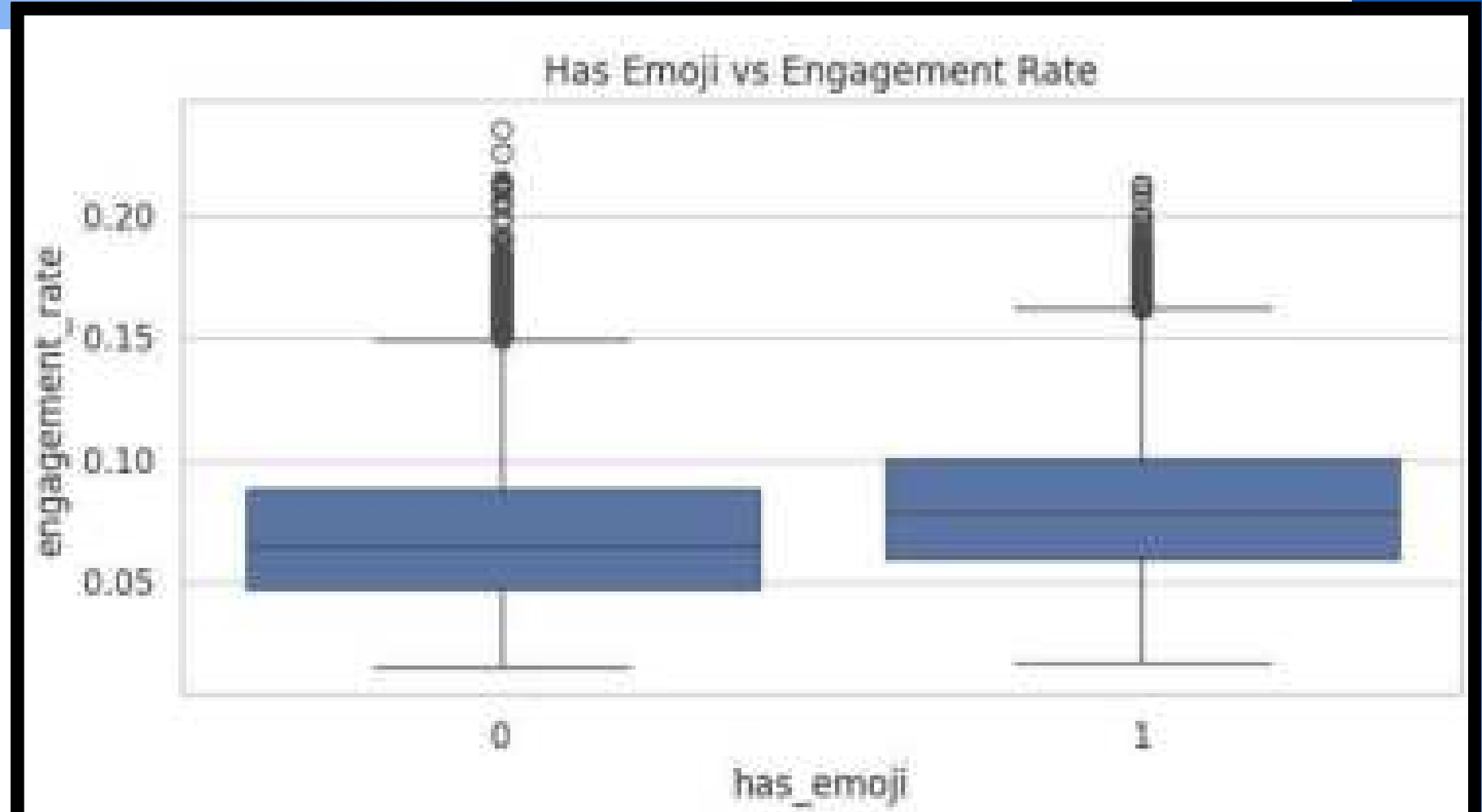


EDA



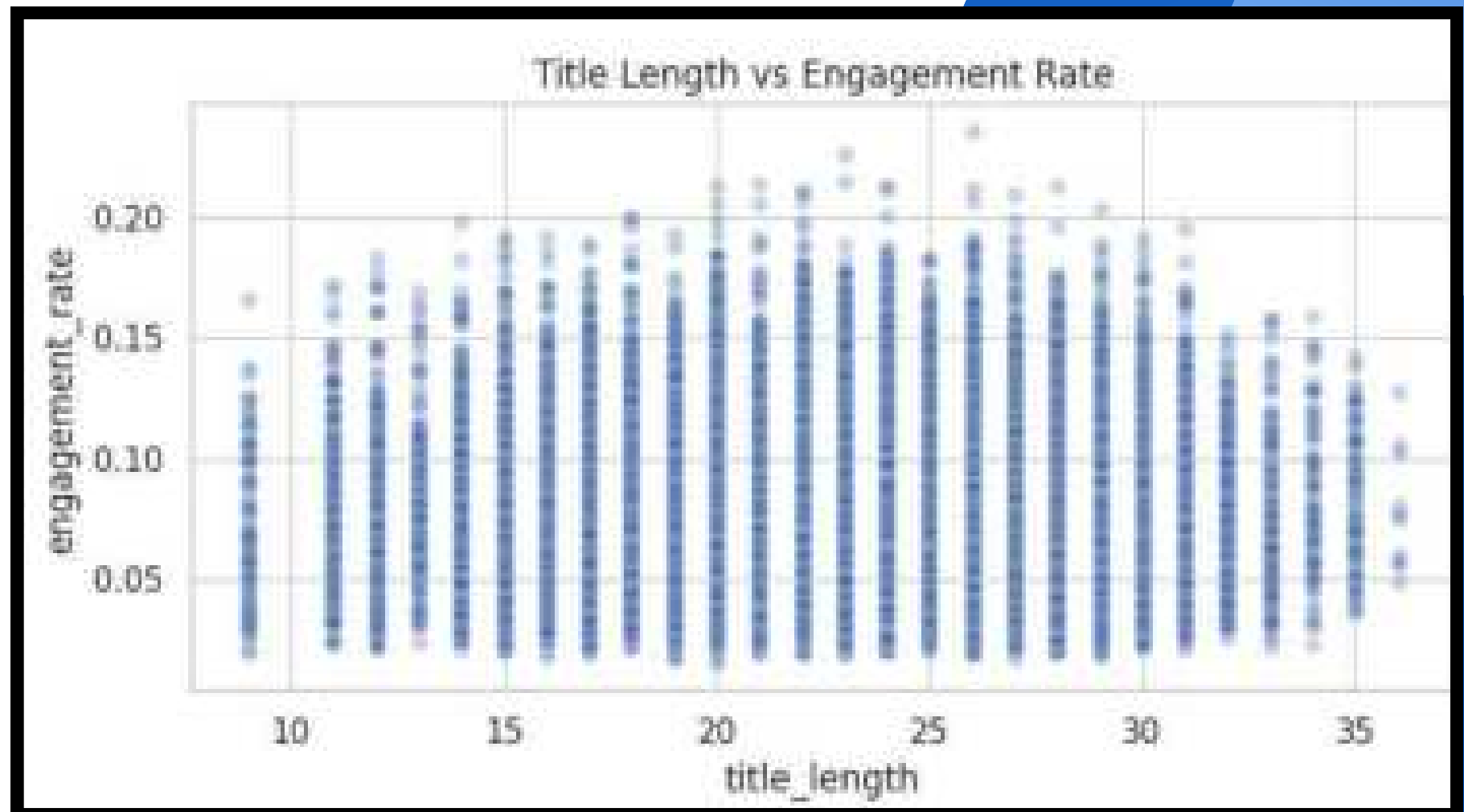
✓ Has Emoji

- Konten dengan emoji (1) memiliki median engagement rate lebih tinggi dibanding tanpa emoji (0)
- Outlier tinggi muncul, tinggi di non tapi padat di grup dengan emoji



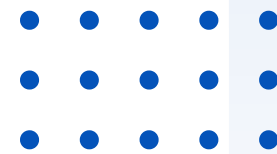
✓ Title length

- Engagement meningkat dari judul pendek menuju menengah
- Puncak engagement 20–25 karakter
- Judul yang terlalu panjang (>30) engagement menurun





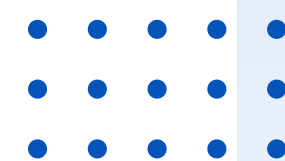
DATA PREPROCESSING



DATA PREPROCESSING

- ✓ **Handling Missing Value**
Data tidak ada missing value
- ✓ **Feature Encoding**
Numerical → Standard Scaler,
Categorical → One Hot Encoder
- ✓ **Data Splitting**
Data Training → 80
Data Testing → 20

MODEL DEVELOPMENT & EVALUATION



Model baseline menghasilkan rata-rata kesalahan prediksi sekitar 2.4 poin engagement rate (MAE \approx 0.024) dengan kesalahan relatif sekitar 33% (sMAPE). Nilai ini menjadi ambang performa minimum yang harus dilampaui oleh model yang dikembangkan selanjutnya agar dapat dianggap memberikan nilai prediktif yang signifikan

BASELINE MODEL

model ini digunakan sebagai tolok ukur minimum yang harus dikalahkan oleh model prediktif selanjutnya

✔ Baseline Performance (Dummy Regressor)

INDICATOR	TRAIN	TEST
MAE	0.02433	0.02437
RMSE	0.03004	0.03031
R SQUARE	0.00000	-0.00027
SMAPE	33.19%	32.99%

MODEL COMPARATION



✓ Linear Regression

cross validation = KFold
(n_splits=3, shuffle=True, random_state=42)

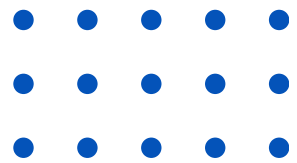
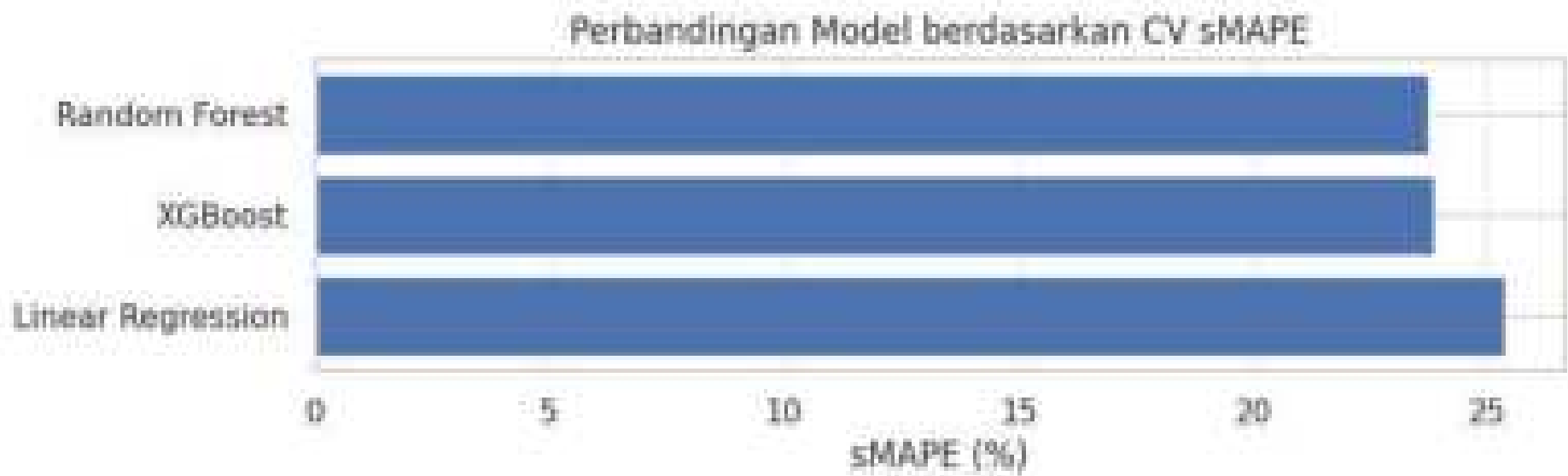
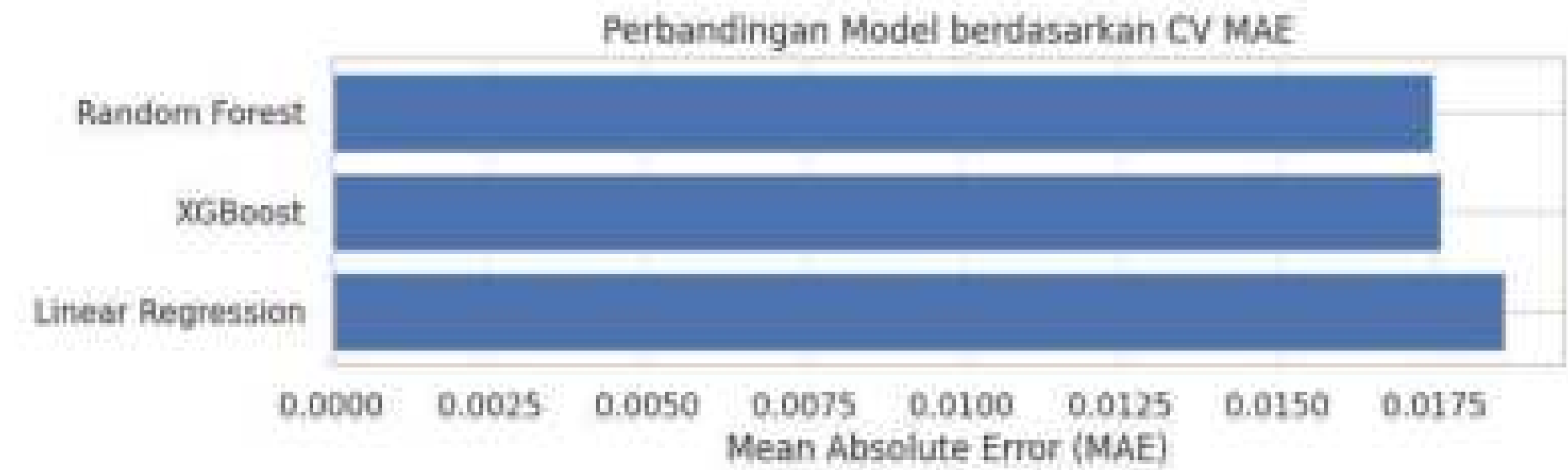
✓ Random Forest

n_estimators = 80,
max_depth = 14,
random_state = 42,
n_jobs = -1

✓ XGBoost

n_estimators = 200,
learning_rate = 0.1,
max_depth = 6,
subsample = 0.8,
colsample_bytree = 0.8,
objective = "reg:squarederror",
random_state = 42,
n_jobs = -1

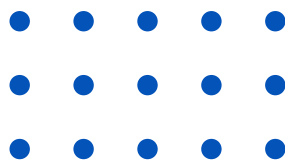
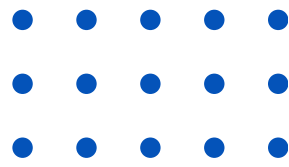
model	CV_MAE	CV_RMSE	CV_R2	CV_sMAPE
Random Forest	0.017478	0.022434	0.442111	23.7302
XGBoost	0.017594	0.022587	0.434505	23.8850
Linear Regression	0.018611	0.023912	0.366192	25.3798



MODEL EVALUATION

FEATURE IMPORTANCE

	feature	importance
1	platform_YouTube	0.515758
2	platform_TikTok	0.458498
3	has_emoji	0.002697
4	creator_avg_views	0.001556
5	duration_sec	0.001177
6	upload_hour	0.000845
7	title_length	0.000698
8	tags_esports, clips, valorant	0.000457
9	tags_beauty, grwm, tips, skincare	0.000449
10	tags_kitchen, recipe, food, fast	0.000388
11	tags_recipe, street food, fast	0.000355
12	tags_lifestyle, routine, motivation	0.000217
13	tags_fps, gaming, clips	0.000209
14	tags_street food, fast, food	0.000194
15	tags_notes, learn, productivity, tips	0.000171



CONCLUSION

Platform adalah Faktor Dominan

platform_YouTube dan platform_TikTok menyumbang mayoritas total importance.
model membedakan engagement berdasarkan platform distribusi, bukan karakteristik konten

Fitur konten pendukung

emoji, duration_sec, creator_avg_views, upload_hour, title_length.
Kontribusinya jauh lebih kecil dibanding platform.
Artinya, efeknya bersifat kondisional setelah platform diketahui.

Fitur Categorical low impact

Tag (tags_*), waktu publikasi (publish_dayofweek_*, publish_period_*), region, hashtag
Importance rendah → pengaruh marginal terhadap prediksi akhir.



MODEL DEPLOYMENT

Model Deployment memungkinkan model menerima data baru dan memberikan output prediksi terhadap End User.
selain itu membantu memonitor performa model dalam dunia nyata.

URL : <https://content-engagement-predictor-app.up.railway.app/>

Content Engagement Predictor

Prediksi engagement rate konten TikTok & YouTube

Platform

TikTok

Creator Tier

Mid (50K views)

Category

Entertainment

Genre

Education

Event Season

Ramadan

Tanggal & Waktu Upload

14 Januari 2026, 21:26

Duration: 60s

Video Title

memasak enak @

Hashtags

#viral

Market Default: Indonesia (id) • Asia • Default Indonesia

Prediksi

Hasil Prediksi (Estimasi)

9% - 10%

1% 20%

Skala 0-20% berdasarkan distribusi data dari pengguna

Rata-rata TikTok: 9.18%

Detail Model dan Filter

View Model Output

model.predict() = 0.092282

Input User:

platform: TikTok
genre: Comedy
category: Entertainment
duration: 60s
tier: Mid
event: Regular

Distribusi Otomatis

title_length: 17
hashtag_count: 1
title_keywords: Jarak Video Viral
hashtag: #viral #viral
upload_hour: 21:00
period: Night
day: Saturday
weekend: Ya
season: Winter
avg_views: 50000

Default: country=id, region=id, language=id, age=18

Prediksi Lagi



Thank You

For your attention to this presentation.