ANALISIS WEBSITE PERFORMANCE DATASET

Ujian Tengah Semester – Data Science

Nama: Peno

NIM: 221220095

Program Studi: Teknik Informatika

Dataset Overview

Aspek	Detail
Sumber	Website Performance Dataset (Kaggle)
Jumlah Data	734 baris (setelah cleaning: 661 baris)
Jumlah Kolom	9 kolom
Variabel Utama	Response Time, Throughput, Load Time, Category, Page Size
Tujuan Analisis	Memahami performa website dan memberikan rekomendasi optimasi

```
1) Jumlah baris dan kolom:
   - Baris : 734
   - Kolom : 9

2) Tipe data setiap kolom:
Sr No                 int64
website_url          object
Category             object
Page Size (KB)      float64
Load Time(s)        float64
Response Time(s)    float64
Throughput          float64
Performance_Label    object
User Response        object
dtype: object

3) Missing values per kolom:
Sr No               0
website_url         0
Category            1
Page Size (KB)      0
Load Time(s)        0
Response Time(s)    0
...
dtype: int64

4) Jumlah baris duplikat:
0
```

Key Metrics:

- **Response Time:** Waktu respons server (detik), indikator kecepatan server.
- **Throughput:** Jumlah request yang diproses per unit waktu, mengukur kapasitas.
- **Load Time:** Waktu muat halaman keseluruhan, pengalaman pengguna akhir.
- **Category:** Jenis konten/channel website, untuk segmentasi performa.

Data Wrangling - Tahap 1: Identifikasi Data

Proses:

- **Identifikasi baris & kolom**: Memahami dimensi awal dataset.
- **Cek tipe data setiap kolom**: Memastikan kesesuaian tipe data untuk analisis.
- **Deteksi missing value**: Mengidentifikasi data yang hilang.
- **Deteksi duplikasi**: Mencari entri data yang berulang.

Hasil:

- **Baris:** 734
- **Kolom:** 9
- **Missing pada Category:** 1 nilai
- **Duplikat:** 0 baris
- **Tipe data:** 5 numerik, 4 kategorikal

Tahap awal ini krusial untuk memastikan kualitas data sebelum analisis lebih lanjut. Mengidentifikasi masalah pada tahap ini membantu mencegah bias dan kesalahan interpretasi di kemudian hari.

Data Wrangling - Tahap 2: Cleaning & Outlier Handling

Bagian A - Imputasi Missing Value:

Kolom Category: Missing 1 nilai → Gunakan **MODE** (kategorikal)

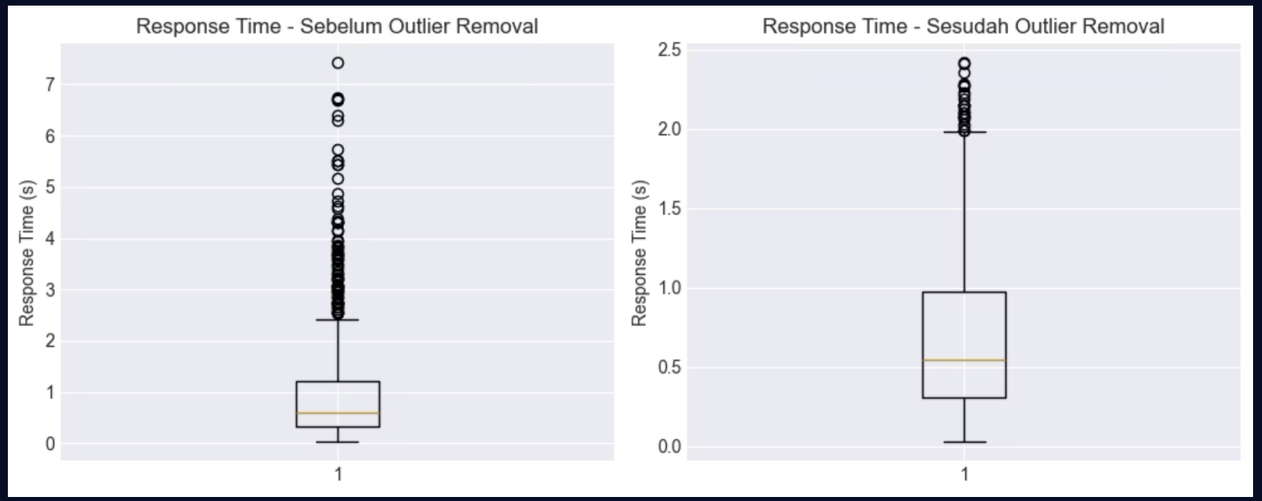
Alasan: Category bersifat kategorikal, mode adalah nilai kategori paling sering yang paling tepat untuk mengisi kekosongan data.

Bagian B - Outlier Detection (IQR Method):

- Q1 = 0.329s, Q3 = 1.202s, IQR = 0.873s
- Lower Bound = -0.980s, Upper Bound = 2.511s
- **Outlier ditemukan:** 72 baris → Dihapus

Bagian C - Hasil Akhir:

- Data sebelum cleaning: 733 baris
- Data setelah cleaning: 661 baris
- Pengurangan: 72 baris (9.8%) dihapus sebagai outlier



Penanganan outlier penting untuk memastikan model tidak bias oleh data yang ekstrem dan tidak representatif.

Statistik Deskriptif - Variabel Numerik Utama

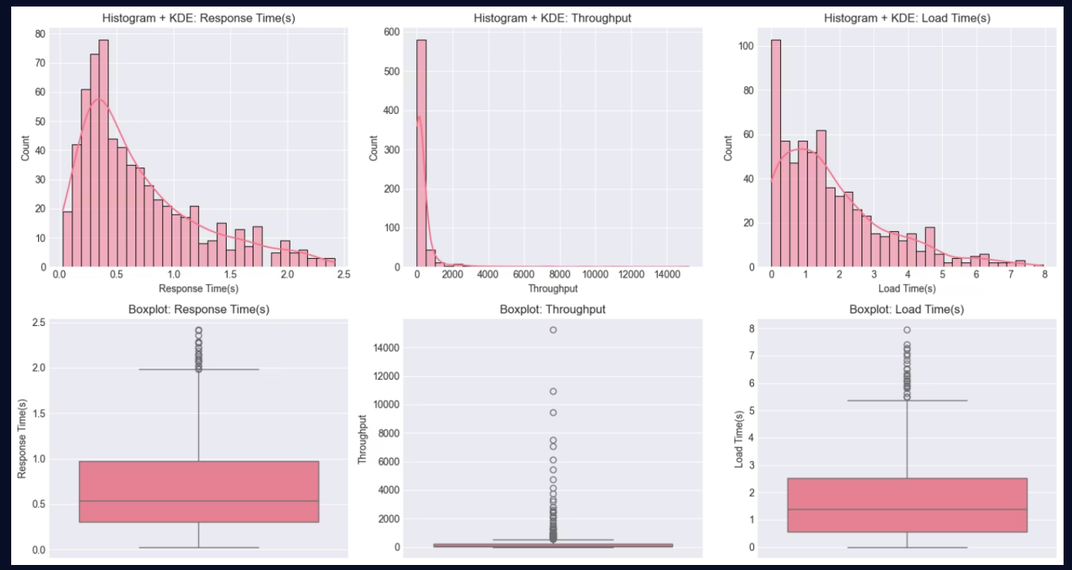
Metrik	Response Time (s)	Throughput	Load Time (s)
Mean	0.704	342.08	1.780
Median	0.542	102.55	1.390
Std Dev	0.528	1057.59	1.585
Min	0.028	0.0	0.0
Max	2.416	15227.28	7.94
Q1	0.304	--	--
Q3	0.975	--	--
P95	1.880	--	--

```
Response Time(s)
  Mean   : 0.704393343419062
  Median: 0.542
  Std    : 0.5284195901254514
  Min    : 0.028
  Max    : 2.416
  Q1     : 0.304
  Q3     : 0.975
  P95    : 1.88

Throughput
  Mean   : 342.0799697428139
  Median: 102.55
  Std    : 1057.5945468446089
  Min    : 0.0
  Max    : 15227.28

Load Time(s)
  Mean   : 1.7795930408472014
  Median: 1.39
  Std    : 1.5853937860546155
  Min    : 0.0
  Max    : 7.94
```

Key Insight:

- **Response Time** cenderung skewed (median < mean), menunjukkan adanya beberapa nilai respons yang lebih lambat.
- **Throughput** memiliki variabilitas sangat tinggi (Std Dev = 1057.59), menandakan fluktuasi besar dalam kapasitas penanganan request.
- **Long tail** terlihat pada P95 = 1.88s (jauh dari median 0.54s) untuk Response Time, mengindikasikan adanya pengguna yang mengalami performa jauh di bawah rata-rata.

Statistik Deskriptif - Visualisasi Distribusi



Histogram + KDE (Response Time, Throughput, Load Time):

Histogram dan Kernel Density Estimate (KDE) memberikan gambaran visual tentang bentuk distribusi data. Kita dapat melihat puncak, penyebaran, dan skewness untuk setiap variabel numerik.

Boxplot (Response Time, Throughput, Load Time):

Boxplot efektif menunjukkan median, kuartil, dan keberadaan outlier. Dari sini, kita bisa memvisualisasikan skewness dan sebaran data secara cepat.

Analisis Distribusi - Skewness & Long Tail

Variabel	Skewness	Interpretasi
Response Time	1.170	Skewed Kanan (Long Tail) ↑ : Distribusi cenderung memiliki 'ekor' panjang ke kanan, artinya ada beberapa respons yang sangat lambat.
Throughput	8.458	Sangat Skewed (Extreme Outlier) ↑↑ : Menunjukkan adanya nilai throughput yang sangat tinggi dan jauh dari rata-rata.
Load Time	1.208	Skewed Kanan (Long Tail) ↑ : Mirip dengan response time, ada beberapa halaman yang membutuhkan waktu muat sangat lama.

```python
for col in num_cols:
    print(col, "skewness =", df_final[col].skew())
```

Python

```
Response Time(s) skewness = 1.1696330231219814
Throughput skewness = 8.457774791453446
Load Time(s) skewness = 1.2078050739756698
```

Implikasi:

1

Mayoritas pengguna

Mayoritas pengguna mengalami response time < 1s (cepat), menunjukkan performa dasar yang baik.

2

Sebagian kecil pengguna

Sebagian kecil pengguna menghadapi response time > 2s (lambat), menciptakan pengalaman yang tidak optimal.

3

Pengalaman tidak seragam

Pengalaman pengguna tidak seragam, membutuhkan perhatian pada segmen yang mengalami kelambatan.

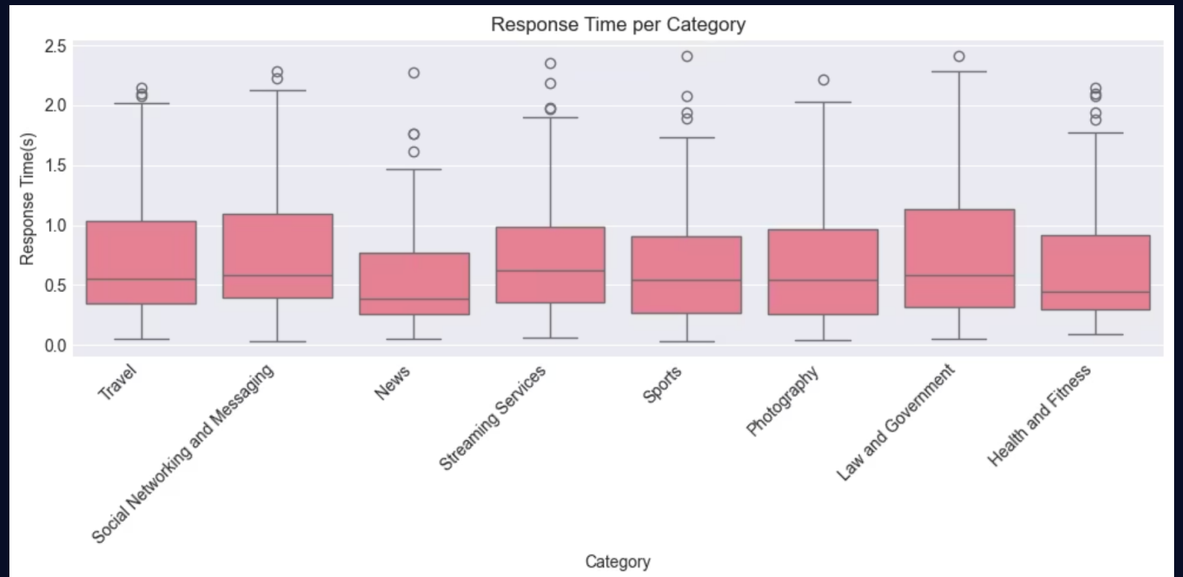
4

Perlu fokus

Perlu fokus pada percentile tinggi (P95, P99) untuk mengidentifikasi dan mengoptimalkan kasus terburuk.

Bivariate Analysis - Response Time per Category

Category	Count	Mean (s)	Median (s)	Std	Min	Max
News	93	0.557	0.389	0.450	0.049	2.270
Health & Fitness	81	0.672	0.448	0.515	0.094	2.146
Sports	87	0.674	0.545	0.517	0.028	2.416
Photography	79	0.702	0.542	0.556	0.041	2.215
Travel	84	0.750	0.554	0.538	0.053	2.146
Streaming Services	100	0.751	0.619	0.504	0.064	2.356
Law & Government	67	0.774	0.586	0.580	0.051	2.414
Social Networking	70	0.791	0.583	0.572	0.036	2.287



Key Finding:

- **Terbaik:** Kategori News dengan rata-rata Response Time 0.557s.
- **Terburuk:** Kategori Social Networking dengan rata-rata Response Time 0.791s.
- **Gap:** Perbedaan 0.234s (~42% lebih lambat), mengindikasikan peluang optimasi spesifik per kategori.

Correlation Analysis - Antar Variabel Numerik

Top 3 Correlations (Absolute):

```
Response Time(s) vs Throughput: r=-0.188
Load Time(s) vs Page Size (KB): r=-0.133
Page Size (KB) vs Throughput: r=-0.059
```

Pasangan Variabel	r	Interpretasi
Response Time vs Throughput	-0.188	Lemah, negatif: Sedikit kecenderungan respons lebih cepat dengan throughput lebih tinggi, namun tidak signifikan.
Load Time vs Page Size	-0.133	Lemah, negatif: Sedikit kecenderungan waktu muat lebih cepat dengan ukuran halaman lebih besar, yang kontraintuitif, mungkin ada faktor lain.
Page Size vs Throughput	-0.059	Sangat lemah: Hampir tidak ada hubungan linear antara ukuran halaman dan throughput.

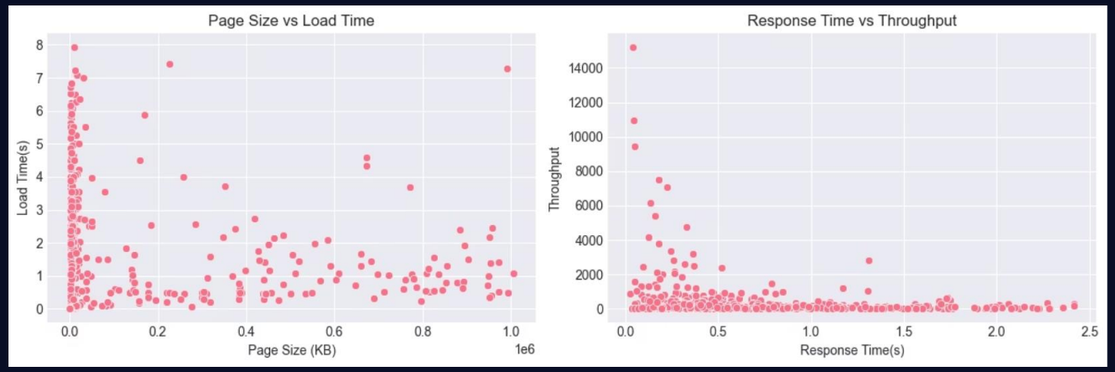
Key Insight:

- Semua korelasi lemah ($|r| < 0.3$), menunjukkan tidak ada hubungan linear yang kuat antar variabel numerik utama.
- Ini mengindikasikan bahwa perlu analisis multi-faktor dan mungkin model yang lebih kompleks (misalnya regresi non-linear atau machine learning) untuk mengidentifikasi pendorong performa utama.

Scatter Plot Analysis - Hubungan Antar Variabel

Interpretasi:

- Pola sebaran, tidak linear kuat
- Variabilitas tinggi, ada banyak outlier
- Hubungan kompleks, bukan one-to-one



Hypothesis Testing - T-Test (Fast vs Slow/Medium)

Data Preparation:

Grup	N	Mean	Median	Std
Fast	254	**0.290s**	0.291s	0.115s
Slow/Medium	407	**0.963s**	0.826s	0.521s

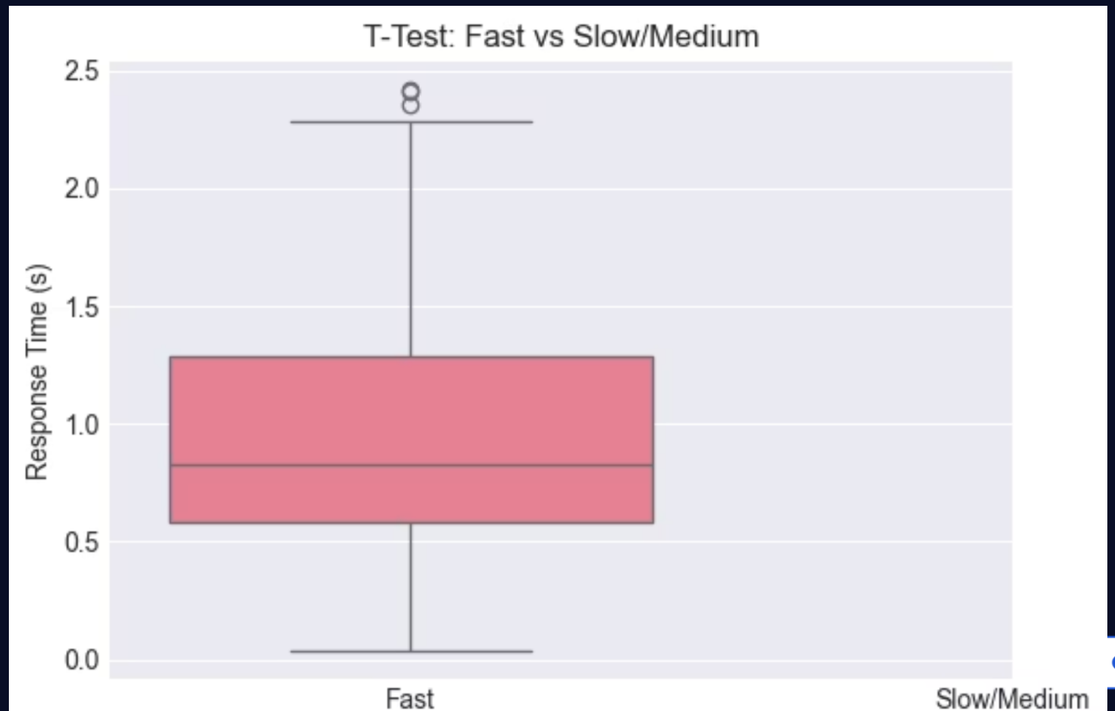
```
Fast: N= 254 mean= 0.2903622047244095 median= 0.2905 std= 0.11548725914302836
Slow/Medium: N= 407 mean= 0.9627813267813268 median= 0.826 std= 0.5209914561530181

T-statistic: -20.25597957521644
p-value    : 2.710739914939783e-71
```

Test Statistics:

- **t-statistic**: -20.256
- **p-value**: 2.71×10^{-71} (< 0.05)
- ✅ **Ada perbedaan signifikan**

Kesimpulan:

- Dua segmen performa benar-benar berbeda
- Fast ~3.3x lebih cepat dari Slow/Medium
- Segmentasi ini valid untuk monitoring



Correlation Analysis - Response Time vs Throughput

Statistik Deskriptif:

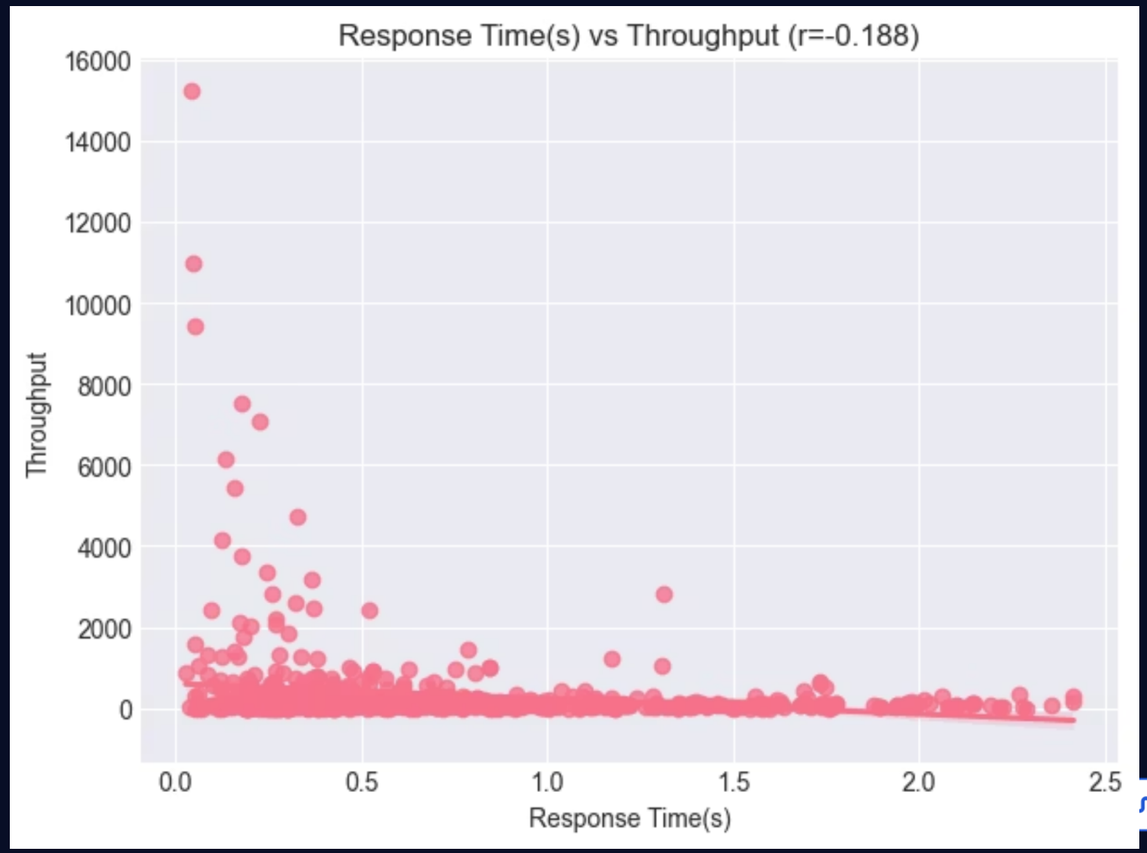
Variabel	Mean	Median	Std
Response Time (s)	0.704	0.542	0.528
Throughput	342.08	102.55	1057.59

Test Result:

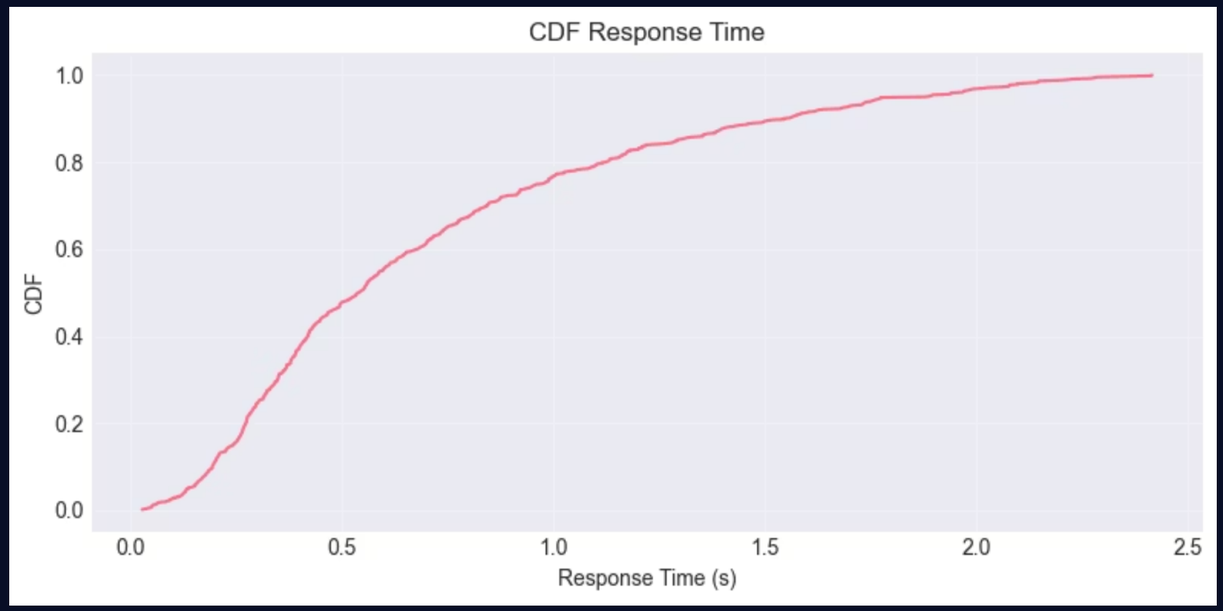
- **Pearson r**: -0.188 (Negatif, Lemah)
- **p-value**: 1.20×10^{-6} (< 0.05, Signifikan)
- ☑ Ada hubungan signifikan, tapi lemah

Interpretasi:

- Korelasi negatif: throughput ↑ → response time ↓ (sedikit)
- Hubungan lemah: banyak variasi lain yang mempengaruhi
- Implikasi: Optimasi tidak bisa hanya fokus throughput



Long Tail Analysis - Performance Distribution



Percentiles & User Impact (kanan):

Percentile	Value	% Pengguna
P50	0.542s	50% ✓ Cepat
P90	1.545s	90% ✓
P95	**1.880s**	**95% ✓**
P99	2.220s	99% ⚠
> 5s	**0%**	✅ Sangat baik

Impact Analysis:

- ✅ 95% pengguna: response < 1.88s
- ⚠ 5% pengguna: response > 1.88s (long tail)
- ❌ Risiko: long tail pada halaman kritikal → bounce, konversi turun
- 💡 Fokus: Optimalkan sisa 5% pengguna di atas P95

Insight Utama - Temuan Analisis (1/2)

Distribusi & User Experience

📊 Mayoritas request cepat (< 1s), tapi ada long tail hingga 2.4s→ Pengalaman pengguna tidak seragam, perlu fokus pada percentile tinggi

Perbedaan Antar Kategori

🥏 Performa bervariasi: News terbaik (0.56s), Social Networking terburuk (0.79s)→ Gap 0.23s (~42%) menunjukkan ruang optimasi per kategori

Segmentasi Performa Signifikan

✅ T-test: Fast (0.29s) vs Slow/Medium (0.96s), p-value < 0.05→ Dua segmen benar-benar berbeda, valid untuk monitoring & prioritas

Korelasi Antar Variabel Lemah

📉 r Response Time vs Throughput = -0.188 (lemah, negatif)→ Tidak ada faktor tunggal dominan, butuh pendekatan multi-faktor

Outlier & Extreme Values

⚠️ Throughput & Load Time sangat skewed, ada sesi dengan beban ekstrem→ Potensi bottleneck di saat-saat beban tinggi

Rekomendasi Strategis - Rencana Aksi (2/2)

Rekomendasi 1: Optimasi Kategori Paling Lambat

🎯 Fokus pada Social Networking & Law/Government (mean 0.77-0.79s)

📋 Aksi: Kompresi asset, optimasi gambar, caching, minifikasi

✅ Target: Kurangi gap 0.2s menuju kategori News (0.56s)

Rekomendasi 2: Mitigasi Long Tail di Atas P95

🔍 Profiling khusus 5% request di atas 1.88s

📋 Aksi: Identifikasi pola (halaman, waktu, kondisi), terapkan fix targeted

✅ Target: Kurangi % di atas P95 dari 5% menjadi 2-3%

Rekomendasi 3: Monitoring Berbasis Label Performa

📊 Gunakan label Fast, Medium, Slow sebagai dashboard KPI

📋 Aksi: Set alert jika % Fast turun atau % Slow naik

✅ Target: Deteksi dini degradasi performa sebelum UX rusak

Rekomendasi 4: Review Arsitektur untuk Throughput Ekstrem

⚙️ Ada sesi dengan throughput 15K+ (vs mean 342), potensi crash

📋 Aksi: Scaling otomatis, load balancing, rate limiting

✅ Target: Jaga stabilitas di beban peak, lindungi user lain

Kesimpulan & Penutup

Summary:

✅ Dataset 734 → 661 baris (clean, outlier removed)✅ Mayoritas pengguna merasakan website responsif (< 1s)✅ Tapi ada long tail 5% di atas P95, perlu mitigasi✅ Perbedaan performa antar kategori signifikan (~42% gap)✅ Korelasi antar variabel lemah → multi-factor approach

Key Takeaway:

🎯 Optimasi performa bukan hanya soal kecepatan rata-rata🎯 Tapi juga mengurangi ekor panjang (long tail) & menjaga konsistensi🎯 Dan fokus pada kategori/sesi dengan performa terburuk

Next Steps:

1. Implementasi 4 rekomendasi strategis
2. Monitor KPI berbasis label (Fast/Medium/Slow)
3. A/B test perubahan optimasi
4. Review quarterly hasil impact