

LAPORAN PROYEK AKHIR

**ANALISIS DATA SEKOLAH DASAR DI DAERAH JAWA TIMUR UNTUK
EVALUASI RASIO GURU-MURID DAN PENYEBARAN SEKOLAH**

Mata Kuliah: Data Wrangling



Disusun oleh:

Atya Azharina Aziz (24031554194)

Muhamad Fadly Saifullah (24031554173)

Kelas: 2024 D

Dosen Pengampu

Dinda Galuh Guminta, M.Stat. (0011129602)

Belgis Ainatul Iza, S.Si., M.Mat. (202509237)

Program Studi S1 Sains Data

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

2024-2025

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Pendidikan merupakan pondasi utama dalam membangun peradaban bangsa. Namun, di Indonesia, kualitas dan pemerataan akses pendidikan dasar masih menjadi tantangan, khususnya di wilayah-wilayah tertentu seperti Jawa Timur yang memiliki karakter geografis dan sosial ekonomi yang beragam. Ketimpangan jumlah sekolah dasar di tiap kabupaten atau kota dapat berdampak pada kesempatan belajar anak-anak, terutama di daerah pedesaan atau tertinggal.

Upaya pemerataan pendidikan ini sejalan dengan Tujuan Pembangunan Berkelanjutan (SDGs) poin ke-4, yaitu *Quality Education*, yang menekankan pentingnya pendidikan berkualitas dan merata bagi semua anak. Selain itu, semangat ini sejalan dengan nilai Asta Cita pertama dan keempat, terutama dalam memperjuangkan hak memperoleh pendidikan dan memperkuat kualitas sumber daya manusia.

Pemerintah Indonesia melalui berbagai kebijakan strategis, seperti program Merdeka Belajar, pemerataan guru, dan peningkatan fasilitas pendidikan, menegaskan komitmen dalam menciptakan sistem pendidikan yang adil dan merata. Proyek ini mendukung arah kebijakan tersebut dengan menggunakan pendekatan data wrangling untuk menganalisis sebaran sekolah dasar di Jawa Timur. Dengan analisis data yang bersih, akurat, dan mudah dipahami, diharapkan hasilnya dapat menjadi dasar bagi perumusan strategi pemerataan pendidikan yang lebih efektif.

1.2 Rumusan Masalah

Analisis ini dibuat untuk menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana sebaran nilai Indeks Pembangunan Manusia (IPM) di Jawa Timur pada tahun 2024?
2. Kabupaten/Kota mana saja yang memiliki nilai IPM tertinggi di Jawa Timur?
3. Apakah ada hubungan visual antara jumlah Sekolah Dasar (SD) dengan nilai IPM di setiap wilayah?

4. Bagaimana perbandingan Angka Partisipasi Kasar (APK) pendidikan di tiap kabupaten/kota terhadap Angka Partisipasi Sekolah (APS) rata-rata provinsi?

1.3. Tujuan Analisis

Berdasarkan rumusan masalah tersebut, tujuan dari analisis ini adalah:

1. Menampilkan distribusi IPM tahun 2024 menggunakan histogram dan boxplot
2. Menyajikan 10 kabupaten/kota dengan IPM tertinggi dalam bentuk diagram batang.
3. Memvisualisasikan potensi hubungan antara jumlah sekolah SD dan IPM menggunakan diagram sebar (*scatter plot*).
4. Membandingkan APK setiap kabupaten/kota dengan nilai APS provinsi menggunakan diagram batang.

1.4 Manfaat Proyek

Pelaksanaan proyek ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Akademik: Mahasiswa dapat memahami bagaimana teknik pengolahan, pembersihan, dan integrasi data digunakan untuk mengidentifikasi pola hubungan antara ketersediaan sekolah dan IPM di Jawa Timur.
2. Manfaat Praktik: Hasil analisis ini dapat menjadi bahan masukan bagi pemerintah daerah maupun instansi pendidikan dalam merumuskan kebijakan pemerataan infrastruktur sekolah dasar di wilayah Jawa Timur.
3. Manfaat Sosial: Masyarakat dapat memperoleh gambaran yang lebih jelas tentang kesenjangan IPM dan ketersediaan infrastruktur pendidikan dasar di provinsi Jawa Timur.
4. Manfaat Teknologis: Proyek ini menunjukkan bagaimana Python dan pustaka sains data dapat digunakan untuk mengolah, membersihkan, dan memvisualisasikan data pendidikan.

BAB 2 TINJAUAN PUSTAKA

2.1 Konsep Pemerataan Pendidikan di Indonesia

Pemerataan pendidikan masih menghadapi tantangan di Jawa Timur, khususnya terkait distribusi guru dan sarana sekolah antara wilayah pedesaan dan perkotaan. Menurut BPS Jawa Timur (2024), ketimpangan fasilitas dan tenaga pendidik masih terlihat di beberapa kabupaten/kota, sehingga pemerintah perlu memahami pola ketimpangan agar kebijakan pendidikan dapat dirumuskan secara lebih akurat.

2.2 Dasar Hukum Pemerataan Pendidikan

Dasar hukum pemerataan pendidikan tercantum dalam Pasal 31 UUD 1945 yang menyatakan bahwa setiap warga negara berhak memperoleh pendidikan dan pemerintah wajib menyelenggarakan pendidikan dasar. Selain itu, Peraturan Pemerintah Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan dan Peraturan Menteri Pendidikan dan Kebudayaan Nomor 23 Tahun 2013 menguraikan upaya peningkatan partisipasi siswa, pemerataan guru, dan penyediaan fasilitas pendidikan di seluruh Indonesia.

2.3 Pemerataan Pendidikan Berdasarkan SDGs

Menurut Perserikatan Bangsa-Bangsa (PBB), Indonesia berkomitmen terhadap Tujuan Pembangunan Berkelanjutan (SDGs), khususnya Tujuan 4: Pendidikan Berkualitas, yang bertujuan memastikan pendidikan yang inklusif dan berkualitas tinggi. SDGs mendorong setiap negara untuk meningkatkan kualitas pendidikan, memperkuat kompetensi guru, serta memperluas akses pembelajaran sepanjang hayat untuk mendukung pembangunan berkelanjutan.

2.4 Asta Cita Nomor 4

Asta Cita Nomor 4 menekankan peningkatan kualitas pendidikan di seluruh Indonesia melalui penguatan karakter, pemerataan infrastruktur pendidikan, peningkatan kompetensi guru, dan digitalisasi pembelajaran (Kementerian Sekretariat Negara RI, 2023). Dalam konteks Jawa Timur, relevansi Asta Cita terlihat pada upaya

pemerintah provinsi meningkatkan kompetensi tenaga pendidik melalui program Guru Penggerak, serta memperluas fasilitas sekolah di wilayah yang masih tertinggal.

2.5 Definisi Data Wrangling

Data wrangling atau data munging adalah proses mengubah, membersihkan, dan menyiapkan data mentah ke dalam format yang siap dianalisis (Kandel et al., 2011). Proses ini mencakup identifikasi kesalahan, pembersihan data, standarisasi format, integrasi data, serta penanganan inkonsistensi agar data dapat digunakan secara optimal.

2.6 Pengolahan Data pada Tahapan Penelitian

Menurut Wickham (2014), komponen utama dari pengolahan data meliputi:

1. Pengumpulan Data, yaitu memperoleh informasi dari berbagai sumber seperti file CSV, web scraping, atau basis data.
2. Pembersihan Data, yaitu menghapus duplikasi, memperbaiki data yang hilang, dan mengatasi inkonsistensi.
3. Integrasi Data, yaitu menggabungkan data dari berbagai sumber agar menjadi satu set data yang utuh.
4. Transformasi Data, yaitu mengubah struktur data, melakukan agregasi, atau mengkonversi format.
5. Visualisasi Data, yaitu menyajikan data dalam bentuk grafik atau diagram untuk membantu pemahaman pola dan tren.

Dalam penelitian ini, pengolahan data dilakukan menggunakan Python dengan memanfaatkan pustaka seperti pandas, numpy, dan matplotlib untuk mengorganisasikan data dari BPS Jawa Timur dan sumber pendidikan resmi lainnya.

BAB 3 METODOLOGI

3.1 Sumber Data, Alat, dan Bahasa Pemrograman

Proyek ini menggunakan metode Data Wrangling sebagai pendekatan utama dalam mempersiapkan dan menganalisis data. Data yang digunakan dalam proyek ini berasal dari tiga sumber, yaitu:

1. Sumber 1 (PDF)

File dari Pusat Data dan Teknologi Informasi Sekretariat Jenderal Kementerian Pendidikan Dasar dan Menengah yang berisi data Indeks Pembangunan Manusia (IPM) dan indikator sosial-ekonomi lainnya, seperti APK, APM, Harapan Lama Sekolah, dan variabel pendidikan lainnya per kabupaten/kota.

2. Sumber 2 (CSV)

Data BPS berjudul Jumlah Sekolah, Guru, dan Murid Sekolah Dasar (SD) di Bawah Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi, berisi jumlah sekolah, guru, dan murid SD (negeri dan swasta) per kabupaten/kota.

3. Sumber 3 (WEB)

Data Angka Partisipasi Sekolah (APS) Provinsi Jawa Timur (nilai 99,34%) yang diekstraksi secara manual dari publikasi daring resmi bernama Harian Disway

4. Sumber 4 (CSV)

Data Angka Partisipasi Sekolah (APS) Provinsi Jawa Timur per kabupaten/kota dari BPS.

Tahapan data wrangling meliputi data collection, data cleaning , data integration , serta data transformation. Visualisasi data kemudian dilakukan menggunakan matplotlib.

Bahasa pemrograman yang digunakan adalah Python dengan bantuan pustaka utama seperti pandas, numpy, dan matplotlib.

3.2 Tahapan Data Wrangling

Tahapan data wrangling pada proyek ini adalah:

1. Data Collection (Pengumpulan Data)

a. Mengambil data dari 4 sumber:

- Sumber 1: Jatim.xlsx data dari hasil scraping file pdf Pusat Data dan Teknologi Informasi Sekretariat Jenderal Kementerian Pendidikan Dasar dan Menengah
- Sumber 2: Jumlah Sekolah, Guru, dan Murid Sekolah Dasar (SD) di Bawah Kementerian Pendidikan, Kebudayaan, Riset.csv. File data BPS yang berisi data kuantitas sekolah, guru, dan murid SD (negeri dan swasta) per kabupaten/kota.
- Sumber 3: Data Angka Partisipasi Sekolah (APS) Provinsi Jawa Timur (99.34%) yang diekstraksi secara manual dari publikasi online.
- Sumber 4 (CSV): Data Angka Partisipasi Sekolah (APS) Provinsi Jawa Timur per kabupaten/kota.

2. Data Cleaning (Pembersihan Data)

a. Memeriksa isi data

```
import pandas as pd
raw_path = r"C:\Users\atyaa\OneDrive\Pictures\MAIN\SEM 3\UAS.Dv\project_wrangle\raw_data"

file_excel = f"{raw_path}\jatim.xlsx"
df_excel = pd.read_excel(file_excel)

file_csv = f"{raw_path}\Jumlah Sekolah, Guru, dan Murid Sekolah Dasar (SD) di Bawah Kementerian Pendidikan, Kebudayaan, Riset,.csv"
df_csv = pd.read_csv(file_csv)

print("Excel Preview:")
display(df_excel.head())

print("\nCSV Preview:")
display(df_csv.head())

Excel Preview:
```

Kabupaten / Kota	Angka Partisipasi Kasar 2024	Angka Partisipasi Murni 2024	Usia Harapan Hidup 2024	Harapan Lama Sekolah 2024	Rata-Rata Lama Sekolah 2024	Pengeluaran Rill per Kapita per Tahun yang Disesuaikan 2024	Indeks Pembangunan Manusia 2024	Kuadra IPM - APM
0 Kab. Pacitan	101,87	95,28	74,74	12,69	7,90	10.099	71,49	
1 Kab. Ponorogo	101,81	96,63	75,28	13,78	7,80	11.065	73,70	
2 Kab. Trenggalek	100,64	96,39	75,35	12,63	7,92	10.872	72,47	
3 Kab. Tulungagung	103,40	99,50	75,20	13,36	8,68	11.966	75,13	
4 Kab. Blitar	100,08	95,91	75,32	12,67	7,87	12.020	73,44	

```
CSV Preview:
```

Kabupaten/Kota	Jumlah Sekolah SD (Negeri)	Jumlah Sekolah SD (Swasta)	Jumlah Sekolah SD (Negeri+Swasta)	Jumlah Guru SD (Negeri)	Jumlah Guru SD (Swasta)	Jumlah Guru SD (Negeri+Swasta)	Jumlah Murid SD (Negeri)	Jumlah Murid SD (Swasta)	Jumlah Murid SD (Negeri+Swasta)
0 Pacitan	408	11	419	3614	161	3775	29126	2777	31903
1 Ponorogo	550	21	571	4031	390	4421	40943	6588	47531
2 Trenggalek	417	21	438	3405	239	3644	35871	3742	39613
3 Tulungagung	569	63	632	5587	810	6397	52132	11076	63208
4 Blitar	634	32	666	5629	371	6000	54552	6581	61133

Data yang ditampilkan berisi indikator pendidikan dan kesejahteraan dari berbagai kabupaten/kota di Indonesia untuk tahun 2024. Beberapa kolom utama dalam dataset meliputi:

- Kabupaten/Kota: Nama wilayah administratif.
- Angka Partisipasi Kasar (APK): Persentase anak usia sekolah yang terdaftar, termasuk yang tidak sesuai usia ideal.
- Angka Partisipasi Murni (APM): Persentase anak usia sekolah yang terdaftar sesuai usia ideal.
- Usia Harapan Hidup: Estimasi rata-rata usia hidup penduduk.
- Harapan Lama Sekolah: Estimasi jumlah tahun pendidikan yang diharapkan ditempuh oleh anak-anak.
- Rata-rata Lama Sekolah: Jumlah tahun pendidikan yang telah ditempuh oleh penduduk usia 15 tahun ke atas.
- Pengeluaran Rill per Kapita per Tahun yang Disesuaikan: Indikator ekonomi yang mencerminkan daya beli dan kesejahteraan masyarakat.

b. Memulai Pre-Processing

```
df_excel.columns = df_excel.columns.str.lower().str.replace(' ', '_')
df_csv.columns = df_csv.columns.str.lower().str.replace(' ', '_')

print(df_excel.columns)
print(df_csv.columns)

Index(['kabupaten/_kota', 'angka_partisipasi_kasar_2024',
       'angka_partisipasi_murni_2024', 'usia_harapan_hidup_2024',
       'harapan_lama_sekolah_2024', 'rata-rata_lama_sekolah_2024',
       'pengeluaran_rill_per_kapita_per_tahun_yang_disesuaikan_2024',
       'indeks_pembangunan_manusia_2024', 'kuadran_ipm_-_apm'],
      dtype='object')
Index(['kabupaten/kota', 'jumlah_sekolah_sd_(negeri)',
       'jumlah_sekolah_sd_(swasta)', 'jumlah_sekolah_sd_(negeri+swasta)',
       'jumlah_guru_sd_(negeri)', 'jumlah_guru_sd_(swasta)',
       'jumlah_guru_sd_(negeri+swasta)', 'jumlah_murid_sd_(negeri)',
       'jumlah_murid_sd_(swasta)', 'jumlah_murid_sd_(negeri+swasta)'],
      dtype='object')
```

Langkah pertama dalam pra-pemrosesan data adalah menstandarisasi nama kolom pada dataset agar konsisten dan mudah diakses. Pada kode ini, semua nama kolom diubah menjadi huruf kecil dan spasi diganti dengan underscore. Hal ini diterapkan pada kedua dataset, yaitu df_excel (data dari Excel) dan df_csv (data dari CSV). Proses ini mempermudah pemanggilan kolom dan menghindari potensi error saat manipulasi data selanjutnya. Nama-nama kolom yang sudah diubah ditampilkan menggunakan fungsi print() untuk memastikan konsistensi.

c. Memeriksa nilai null/NaN

```
print("Missing values in Excel:")
print(df_excel.isna().sum())

print("\nMissing values in CSV:")
print(df_csv.isna().sum())

Missing values in Excel:
kabupaten/_kota                                0
angka_partisipasi_kasar_2024                  0
angka_partisipasi_murni_2024                  0
usia_harapan_hidup_2024                      0
harapan_lama_sekolah_2024                     0
rata-rata_lama_sekolah_2024                  0
pengeluaran_rill_per_kapita_per_tahun.yang_disesuaikan_2024 0
indeks_pembangunan_manusia_2024              0
kuadran_ipm_-_apm                            0
dtype: int64

Missing values in CSV:
kabupaten/kota                                0
jumlah_sekolah_sd_(negeri)                   0
jumlah_sekolah_sd_(swasta)                   0
jumlah_sekolah_sd_(negeri+swasta)             0
jumlah_guru_sd_(negeri)                      0
jumlah_guru_sd_(swasta)                      0
jumlah_guru_sd_(negeri+swasta)                0
jumlah_murid_sd_(negeri)                     0
jumlah_murid_sd_(swasta)                     0
jumlah_murid_sd_(negeri+swasta)               0
dtype: int64
```

Langkah ini merupakan bagian dari data cleaning, yaitu proses validasi kelengkapan data sebelum dilakukan analisis lebih lanjut. Dengan memastikan tidak ada nilai kosong, pengguna dapat melanjutkan ke tahap analisis statistik, visualisasi, atau pemodelan tanpa risiko error akibat data yang tidak lengkap

d. Memeriksa data duplikat (Ditemukan 0)

```
print("Duplicate rows in Excel:", df_excel.duplicated().sum())
print("Duplicate rows in CSV:", df_csv.duplicated().sum())

Duplicate rows in Excel: 0
Duplicate rows in CSV: 0
```

yaitu validasi integritas data dengan memastikan tidak ada baris yang terduplicasi. Duplikasi data dapat menyebabkan bias dalam analisis statistik, perhitungan agregat yang tidak akurat, dan hasil model yang tidak valid. Dengan hasil nol duplikasi

- e. Mengkonversi format angka (menggunakan koma) ke tipe data numerik (float)

```
for col in df_excel.columns:  
    df_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
  
print(df_excel.dtypes)  
  
Python  
  
paten_`_kota          object  
a_partisipasi_kasar_2024      float64  
a_partisipasi_murni_2024      float64  
_harapan_hidup_2024         float64  
pan_lama_sekolah_2024        float64  
-rata_lama_sekolah_2024      float64  
eluaran_rill_per_kapita_per_tahun_yang_disesuaikan_2024  float64  
ks_pembangunan_manusia_2024  float64  
ran_ipm_`_apm              object  
e: object  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')  
seri\atya\AppData\Local\Temp\ipykernel_8108\904512835.py:2: FutureWarning: errors='ignore' is deprecated and will raise in a future version. Use to_num  
_excel[col] = pd.to_numeric(df_excel[col].astype(str).str.replace(',', '.'), errors='ignore')
```

- Menstandarkan format angka agar sesuai dengan sistem desimal mengganti koma dengan titik.
- Mengubah tipe data menjadi numerik agar bisa digunakan dalam analisis statistik, visualisasi, atau pemodelan.
- Menghindari error saat melakukan operasi matematis atau agregasi data.

- f. Mengidentifikasi dan menangani *outlier* menggunakan metode IQR dan teknik *Winsorization*

```

import pandas as pd

def detect_and_display_outliers(df):
    numeric_cols = df.select_dtypes(include='number').columns
    outlier_reports = []

    for col in numeric_cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower = Q1 - 1.5 * IQR
        upper = Q3 + 1.5 * IQR

        outliers = df[(df[col] < lower) | (df[col] > upper)]

        if not outliers.empty:
            for idx, row in outliers.iterrows():
                outlier_reports.append({
                    'column': col,
                    'index': idx,
                    'value': row[col],
                    'lower_bound': lower,
                    'upper_bound': upper,
                    'kabupaten_kota': row.get('kabupaten_kota')
                })

    if outlier_reports:
        return pd.DataFrame(outlier_reports)
    else:
        return "Tidak ada outlier yang terdeteksi di kolom numerik."

outlier_summary = detect_and_display_outliers(df_excel)

```

	column	index	value	lower_bound	upper_bound	kabupaten_kota
0	angka_partisipasi_kasar_2024	25	107.740	96.482500	107.382500	Kab. Bangkalan
1	angka_partisipasi_kasar_2024	26	110.700	96.482500	107.382500	Kab. Sampang
2	angka_partisipasi_kasar_2024	27	108.240	96.482500	107.382500	Kab. Pamekasan
3	angka_partisipasi_kasar_2024	28	108.020	96.482500	107.382500	Kab. Sumenep
4	angka_partisipasi_kasar_2024	29	113.140	96.482500	107.382500	Kota Kediri
5	harapan_lama_sekolah_2024	29	15.710	11.261250	15.671250	Kota Kediri
6	harapan_lama_sekolah_2024	31	15.790	11.261250	15.671250	Kota Malang
7	pengeluaran_rill_per_kapita_per_tahun_yang_dis...	36	19.666	7.216625	17.821625	Kota Surabaya

```

processed_path = r"C:\Users\atyaa\OneDrive\Pictures\MAIN\SEM 3\UAS.DW\project_wrangle\processed_data"

df_excel.to_csv(f"{processed_path}\jatim_clean_step1.csv", index=False)
df_csv.to_csv(f"{processed_path}\sd_clean_step1.csv", index=False)

```

```

def missing_info(df):
    missing = df.isna().sum()
    percent = (df.isna().sum() / len(df)) * 100
    return pd.DataFrame({'missing': missing, 'percent': percent})

```

```

print("Excel:")
print(missing_info(df_excel))
print("\nCSV:")
print(missing_info(df_csv))

```

Excel:

	missing	percent
kabupaten/_kota	0	0.0
angka_partisipasi_kasar_2024	0	0.0
angka_partisipasi_murni_2024	0	0.0
usia_harapan_hidup_2024	0	0.0
harapan_lama_sekolah_2024	0	0.0
rata-rata_lama_sekolah_2024	0	0.0
pengeluaran_rill_per_kapita_per_tahun_yang_dise...	0	0.0
indeks_pembangunan_manusia_2024	0	0.0
kuadran_ipm_-_apm	0	0.0

CSV:

	missing	percent
kabupaten/kota	0	0.0
jumlah_sekolah_sd_(negeri)	0	0.0
jumlah_sekolah_sd_(swasta)	0	0.0
jumlah_sekolah_sd_(negeri+swasta)	0	0.0
jumlah_guru_sd_(negeri)	0	0.0
jumlah_guru_sd_(swasta)	0	0.0
jumlah_guru_sd_(negeri+swasta)	0	0.0
jumlah_murid_sd_(negeri)	0	0.0
jumlah_murid_sd_(swasta)	0	0.0
jumlah_murid_sd_(negeri+swasta)	0	0.0

```

print("Kolom di df_excel:", df_excel.columns.tolist())
print("Kolom di df_csv:", df_csv.columns.tolist())

```

Kolom di df_excel: ['kabupaten/_kota', 'angka_partisipasi_kasar_2024', 'angka_partisipasi_murni_2024', 'usia_harapan_hidup_2024', 'harapan_lama_sekolah_2024', 'rata-rata_lama_sekolah_2024', 'pengeluaran_rill_per_kapita_per_tahun_yang_dise...', 'indeks_pembangunan_manusia_2024', 'kuadran_ipm_-_apm']
Kolom di df_csv: ['kabupaten/kota', 'jumlah_sekolah_sd_(negeri)', 'jumlah_sekolah_sd_(swasta)', 'jumlah_sekolah_sd_(negeri+swasta)', 'jumlah_guru_sd_(negeri)', 'jumlah_guru_sd_(swasta)', 'jumlah_guru_sd_(negeri+swasta)', 'jumlah_murid_sd_(negeri)', 'jumlah_murid_sd_(swasta)', 'jumlah_murid_sd_(negeri+swasta)']

- Mengidentifikasi nilai ekstrem yang dapat mempengaruhi analisis.
- Menentukan apakah perlu dilakukan penanganan seperti penghapusan, imputasi, atau transformasi data.
- Menjaga kualitas dan keandalan hasil analisis statistik atau pemodelan.
- Dalam penelitian ini seluruh outlier tidak dihapus dikarenakan untuk mendapatkan data asli (*real*).

3.3 Data Integration (Integrasi Data)

Tahapan data wrangling pada proyek ini adalah:

1. Membuat kolom kunci norm_name untuk standarisasi nama wilayah

```
def simple_normalize(name):
    name = str(name).strip()
    if name.startswith('Kota'):
        return name.replace('Kota', '').strip() + '_Kota'
    elif name.startswith('Kab.'):
        return name.replace('Kab.', '').strip() + '_Kab'
    else:
        return name + '_Kab' # default ke Kabupaten

# Apply ke kedua dataset
df_ipm_clean['norm_name'] = df_ipm_clean['kabupaten/_kota'].apply(simple_normalize)
df_sd_clean['norm_name'] = df_sd_clean['kabupaten/kota'].apply(simple_normalize)

# Cek hasil
print("Sample normalized names:")
print(df_ipm_clean[['kabupaten/_kota', 'norm_name']].head(40))
print(df_sd_clean[['kabupaten/kota', 'norm_name']].head(40))
```

```
Sample normalized names:
      kabupaten/_kota      norm_name
0      Kab. Pacitan      Pacitan_Kab
1      Kab. Ponorogo     Ponorogo_Kab
2      Kab. Trenggalek   Trenggalek_Kab
3      Kab. Tulungagung  Tulungagung_Kab
4      Kab. Blitar       Blitar_Kab
5      Kab. Kediri       Kediri_Kab
6      Kab. Malang       Malang_Kab
7      Kab. Lumajang     Lumajang_Kab
8      Kab. Jember       Jember_Kab
9      Kab. Banyuwangi   Banyuwangi_Kab
10     Kab. Bondowoso   Bondowoso_Kab
11     Kab. Situbondo   Situbondo_Kab
12     Kab. Probolinggo  Probolinggo_Kab
13     Kab. Pasuruan    Pasuruan_Kab
14     Kab. Sidoarjo    Sidoarjo_Kab
15     Kab. Mojokerto   Mojokerto_Kab
16     Kab. Jombang     Jombang_Kab
17     Kab. Nganjuk     Nganjuk_Kab
18     Kab. Madiun      Madiun_Kab
19     Kab. Magetan     Magetan_Kab
20     Kab. Ngawi       Ngawi_Kab
21     Kab. Bojonegoro  Bojonegoro_Kab
22     Kab. Tuban       Tuban_Kab
...
35     Kota Madiun     Madiun_Kota
36     Kota Surabaya   Surabaya_Kota
37     Kota Batu       Batu_Kota
38     Jawa Timur      Jawa Timur_Kab
```

- a. Menyatukan format penamaan wilayah agar konsisten.
- b. Mempermudah proses penggabungan antar dataset berdasarkan wilayah.

- c. Menghindari kesalahan pencocokan data akibat perbedaan penulisan nama kabupaten/kota.
2. Menggabungkan (merge) dua data bersih (Data 1 dan Data 2) menjadi satu data utama

```

# Merge
df_final = pd.merge(df_ipm_clean, df_sd_clean, on='norm_name', how='inner')

print(f"Final shape: {df_final.shape}")
print(f"Duplicates: {df_final['norm_name'].duplicated().sum()}")


Final shape: (38, 28)
Duplicates: 0


# Hapus kolom redundant
columns_to_drop = [
    'nama_normalized',
    'nama_normalized_v2',
    'norm_name'
]

df_final_clean = df_final.drop(columns=[col for col in columns_to_drop if col in df_final.columns])
print(f"Shape setelah cleanup: {df_final_clean.shape}")


# Save ke file processed
output_path = r"C:\Users\atyaa\OneDrive\Pictures\MAIN\SEM 3\UAS.DW\project_wrangle\processed_data"
df_final_clean.to_csv(f"{output_path}\jatim_merged_final.csv", index=False)
print("Data berhasil disimpan!")


Data berhasil disimpan!

# Final validation
print("FINAL VALIDATION")
print(f"Total districts: {len(df_final_clean)}")
print(f"Missing values: {df_final_clean.isnull().sum().sum()}")
print(f"Duplicate districts: {df_final_clean['kabupaten/_kota'].duplicated().sum()}")


print("\nSample final data:")
print(df_final_clean[['kabupaten/_kota', 'indeks_pembangunan_manusia_2024', 'jumlah_sekolah_sd_(negeri+swasta)']].head())


FINAL VALIDATION
Total districts: 38
Missing values: 0
Duplicate districts: 0

Sample final data:
  kabupaten/_kota indeks_pembangunan_manusia_2024 \
0   Kab. Pacitan           71.49
1   Kab. Ponorogo          73.70
2   Kab. Trenggalek         72.47
3   Kab. Tulungagung        75.13
4   Kab. Blitar              73.44

  jumlah_sekolah_sd_(negeri+swasta)
0                           419
1                           571
2                           438
3                           632
4                           666

```

- a. Menghapus kolom redundant
- b. Menyimpan hasil merged
- c. Memeriksa ulang hasil validasi akhir

3.4 Data Analysis & Exploration (Analisis dan Eksplorasi)

Langkah eksplorasi dilakukan untuk memahami karakteristik data. Prosesnya meliputi:

- a. Distribusi dan Pemeriksaan Variabel
 - Pemeriksaan distribusi IPM (histogram, boxplot).
 - Pemeriksaan nilai partisipasi pendidikan (APK, APM, APS).
 - Pemeriksaan distribusi jumlah sekolah, guru, dan murid.
 - Pemeriksaan rasio murid per guru.
- b. Analisis Perbandingan dan Hubungan Antar Variabel
 - Analisis Top 10 IPM kabupaten/kota.
 - Scatter plot hubungan:
 - Jumlah sekolah vs IPM
 - Jumlah guru vs IPM
 - Jumlah murid vs IPM
 - Rasio murid-guru vs IPM
 - APK vs rasio murid-guru
 - APM vs rasio murid-guru
 - HLS vs rasio murid-guru
- c. Analisis APK wilayah vs APS Provinsi.
- d. Heatmap korelasi antara:
 - APK, APM, IPM
 - Jumlah sekolah, guru, murid
 - Rasio murid-guru

3.5 Data Visualization & Publishing (Visualisasi dan Publikasi)

Tahap visualisasi dilakukan untuk menyajikan hasil analisis secara informatif dan mudah dipahami. Kegiatan yang dilakukan meliputi:

1. Membuat berbagai jenis visualisasi
Termasuk:
 - a. Scatter plot untuk melihat hubungan dua variabel.

- b. Histogram untuk melihat distribusi data,
 - c. Boxplot untuk melihat persebaran dan outlier,
 - d. Heatmap korelasi untuk memetakan kekuatan hubungan antar variabel
2. Memperbaiki estetika visualisasi
Seperti menambahkan judul, label, ukuran figur, serta pemilihan warna yang konsisten dengan standar visual laporan.
3. Mempublikasikan data hasil wrangling
Menyimpan dataset gabungan yang telah dibersihkan dalam file jatim_merged_final.csv, yang kemudian digunakan sebagai dasar analisis visual, korelasi, dan pembahasan pada bab selanjutnya.
4. Menyiapkan visualisasi untuk laporan akhir
Semua grafik yang relevan dimasukkan ke dalam laporan untuk mendukung pembahasan pada Bab Hasil dan Pembahasan.

BAB 4 HASIL DAN PEMBAHASAN

Bab ini membahas temuan yang diperoleh dari proses data wrangling dan analisis eksplorasi data (EDA) yang dilakukan dalam notebook.

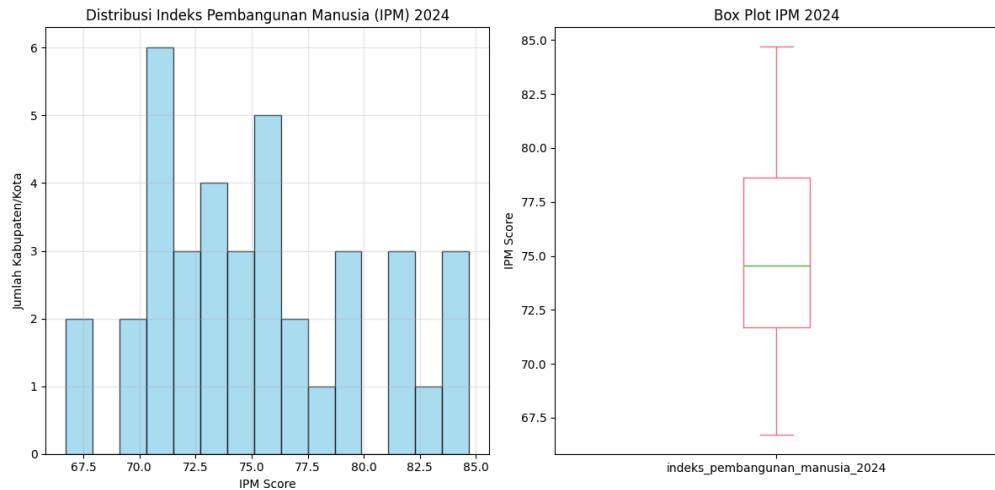
4.1 Hasil Pembersihan dan Integrasi Data

Proses data wrangling dimulai dengan pembersihan dan persiapan dua dataset utama

1. Hasil Pembersihan dan Integrasi Data:
 - a. Data berhasil digabung menjadi dataset final berisi 38 kabupaten/kota.
 - b. Tipe data numerik berhasil dikonversi.
 - c. Outlier pada jumlah murid dan rasio murid-guru teridentifikasi di beberapa wilayah.
 - d. Dataset final siap untuk dianalisis.

4.2 Analisis Data dan Visualisasi

Analisis eksplorasi data (EDA) dari data yang telah dibersihkan memberikan wawasan berikut:



Langkah ini bertujuan untuk menyajikan hasil analisis data secara visual agar lebih mudah dipahami dan diinterpretasikan. Beberapa jenis visualisasi yang digunakan meliputi:

1. Histogram

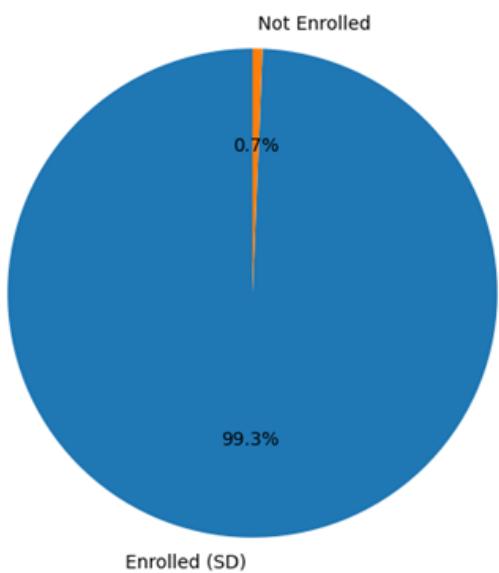
Digunakan untuk menampilkan distribusi nilai dari variabel kuantitatif seperti Indeks Pembangunan Manusia (IPM), pengeluaran per kapita, atau rata-rata lama sekolah. Histogram membantu mengidentifikasi pola sebaran data, seperti apakah data bersifat normal, miring (skewed), atau memiliki outlier.

2. Boxplot

Digunakan untuk mendeteksi nilai pencilan dan memahami sebaran data berdasarkan kuartil. Visualisasi ini sangat berguna dalam membandingkan indikator antar wilayah atau kelompok.

3. Tingkat Partisipasi Pendidikan

Angka Partisipasi Sekolah (APS) Jenjang SD di Jawa Timur (2025)

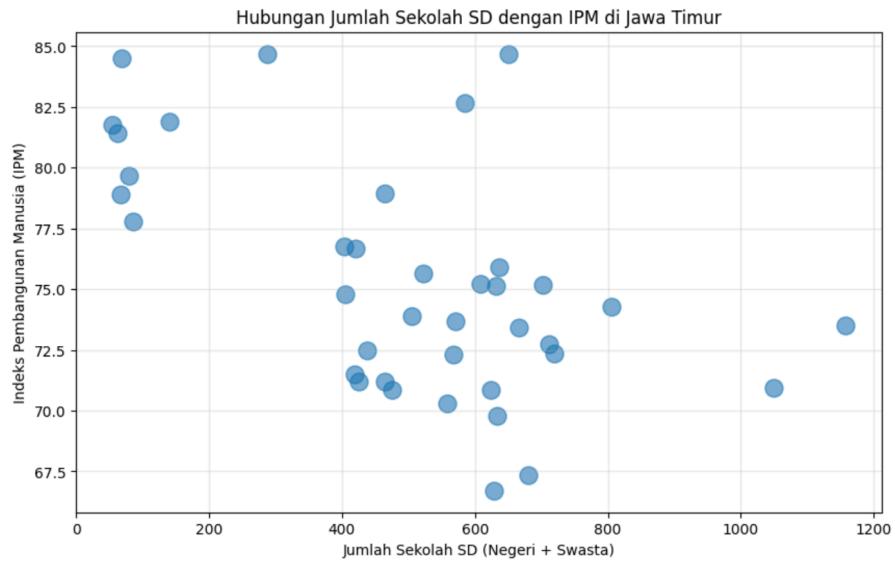


Data ketiga (statistik APS Provinsi) menunjukkan Angka Partisipasi Sekolah (APS) SD di Jawa Timur adalah 99.34%. Angka ini divisualisasikan dalam Pie Chart menunjukkan bahwa partisipasi dan akses pendidikan dasar di Jatim secara umum sudah sangat baik dan hampir sempurna (merata).

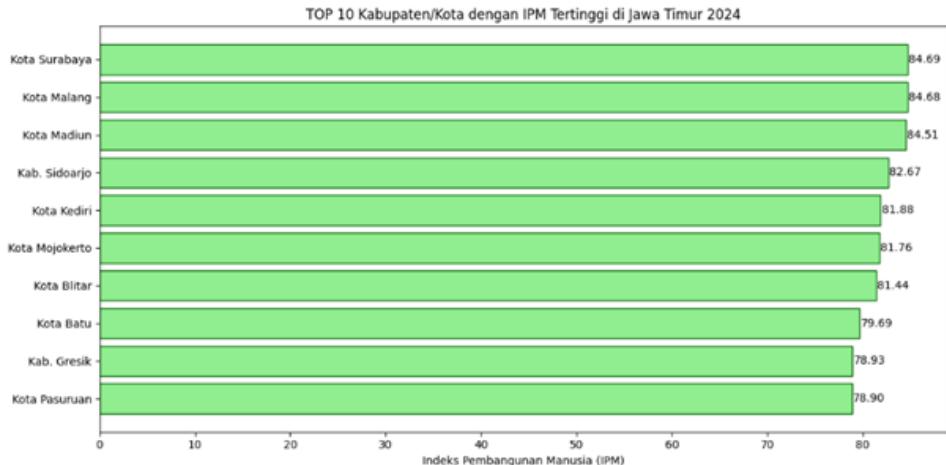
4. Kesenjangan IPM

Meskipun partisipasi sekolah merata, kondisi pembangunan manusia (IPM) menunjukkan kesenjangan.

- Distribusi IPM: Histogram dan Boxplot menunjukkan bahwa nilai IPM di Jatim bervariasi, dengan sebaran data yang cukup luas dan beberapa nilai tinggi yang menonjol.

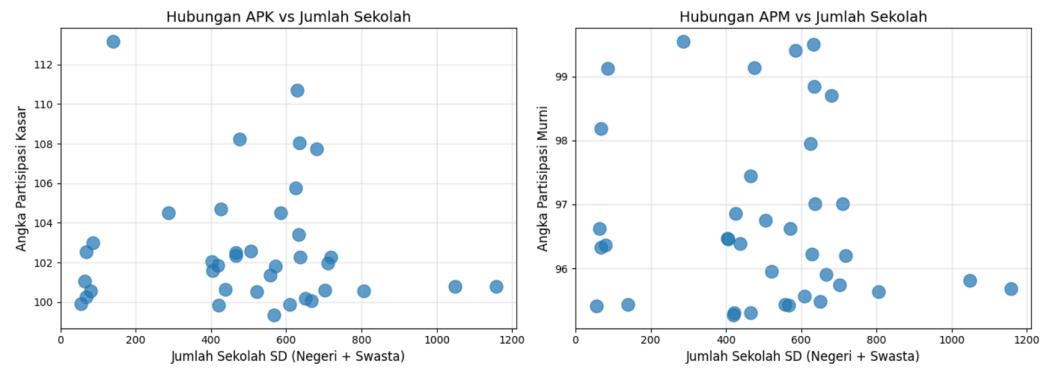


- Top 10 IPM:



Grafik batang menunjukkan bahwa 10 besar IPM didominasi oleh wilayah perkotaan (Kota Surabaya, Kota Malang, Kota Madiun, dll.), yang mengindikasikan adanya kesenjangan IPM yang signifikan antara wilayah urban dan rural di Jawa Timur.

5. Hubungan APK dan APM vs Jumlah sekolah



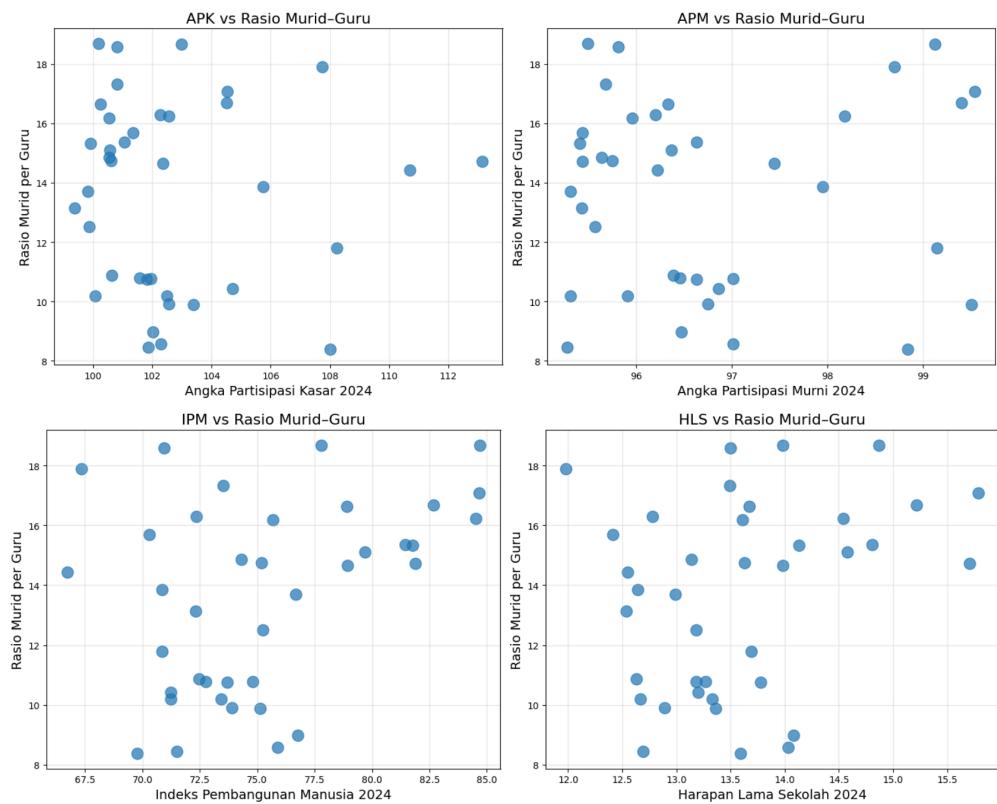
a. Plot kiri: Hubungan antara Jumlah Sekolah SD dan Angka Partisipasi Kasar (APK)

- Titik-titik menunjukkan variasi APK pada setiap kabupaten/kota.
- Polanya menyebar, sehingga hubungan terlihat lemah atau tidak jelas.

b. Plot kanan: Hubungan antara Jumlah Sekolah SD dan Angka Partisipasi Murni (APM). Titik-titik juga tersebar acak, menunjukkan hubungan yang kurang kuat.

Secara umum, kedua grafik menunjukkan bahwa jumlah sekolah tidak memiliki hubungan langsung yang kuat dengan APK maupun APM.

6. Hubungan APK, APM, IPM, HLS vs Rasio Murid-Guru



a. APK vs Rasio Murid–Guru

Angka Partisipasi Kasar (APK) cenderung stabil dan tinggi (berkisar 100–112%) meskipun rasio murid per guru bervariasi dari 8 hingga 16. Hal ini menunjukkan bahwa rasio guru yang lebih rendah tidak secara signifikan menghambat akses pendidikan secara kasar. Stabilitas ini mencerminkan keberhasilan program pendidikan dasar dalam menjangkau populasi usia sekolah, bahkan di daerah dengan keterbatasan tenaga pendidik.

b. APM vs Rasio Murid–Guru

Nilai IPM cenderung menurun, terutama di atas rasio 12. Ini menunjukkan bahwa ketersediaan guru yang memadai berkaitan erat dengan peningkatan kualitas hidup secara menyeluruh (pendidikan, kesehatan, ekonomi), bukan hanya akses sekolah.

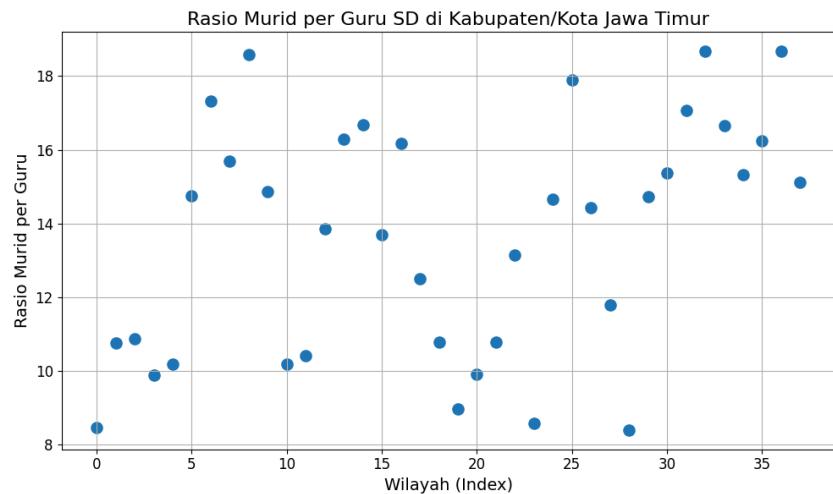
c. IPM vs Rasio Murid–Guru

IPM juga relatif tinggi dan tidak menunjukkan penurunan yang signifikan seiring kenaikan rasio murid per guru. Hal ini menegaskan bahwa partisipasi murni peserta didik pada jenjang yang tepat lebih dipengaruhi oleh faktor kebijakan dan sosial ekonomi daripada rasio guru-murid.

d. HLS vs Rasio Murid–Guru

Daerah dengan rasio murid per guru yang tinggi, harapan lama sekolah cenderung lebih rendah karena keterbatasan kualitas dan daya dukung pembelajaran.

7. Hubungan Rasio Murid per Guru SD di kabupaten jawa timur



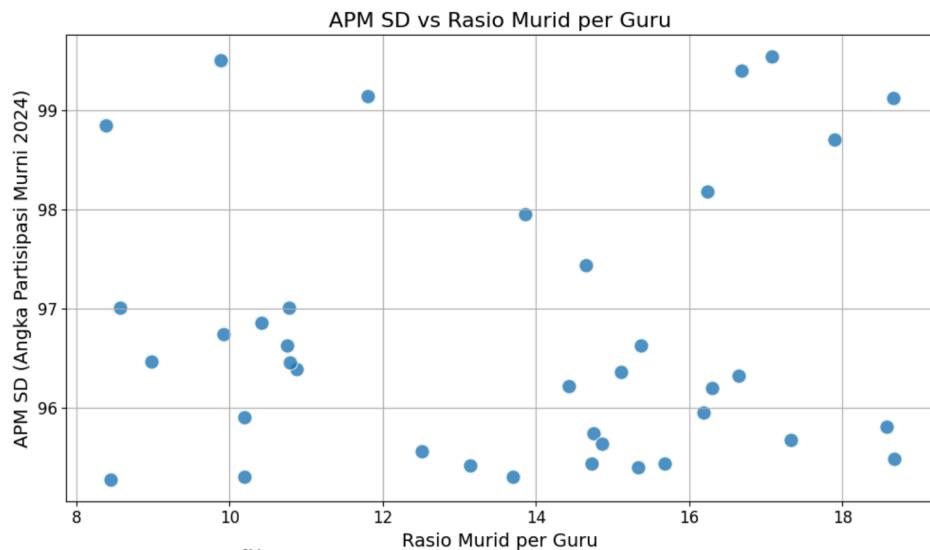
a. Komponen Grafik:

- Sumbu X (Horizontal): Melabelkan "Wilayah (Index)", yang mewakili urutan atau kode dari berbagai Kabupaten/Kota (dari 0 hingga 38).
- Sumbu Y (Vertikal): Melabelkan "Rasio Murid per Guru", yang menunjukkan angka perbandingan jumlah murid terhadap satu orang guru.
- Titik Data: Titik-titik biru menunjukkan nilai rasio untuk setiap wilayah.

b. Interpretasi Data:

- Nilai rasio terendah berada di kisaran angka 8, sedangkan nilai rasio tertinggi mendekati angka 19.
- Sebaran ini menunjukkan adanya ketimpangan beban kerja guru antar wilayah di Jawa Timur; beberapa daerah memiliki sedikit murid per guru, sementara daerah lain memiliki hampir 19 murid untuk satu guru.

8. Hubungan APM vs Rasio Murid per Guru SD di kabupaten jawa timur



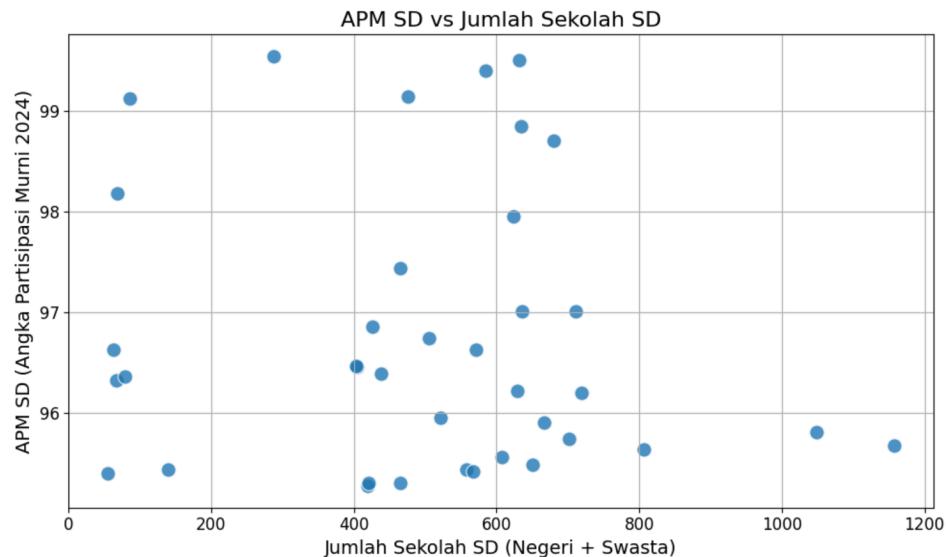
a. Komponen Grafik:

- Sumbu X (Horizontal): Mewakili "Rasio Murid per Guru", dengan rentang angka dari sekitar 8 hingga 19.
- Sumbu Y (Vertikal): Mewakili "APM SD (Angka Partisipasi Murni 2024)", dengan skala persentase dari sekitar 95 hingga hampir 100.

b. Interpretasi Data:

- Sebaran Acak: Titik-titik data tersebar cukup acak di seluruh area grafik tanpa membentuk pola garis yang jelas (baik naik maupun turun).
- Kesimpulan Awal: Pola ini mengindikasikan tidak adanya korelasi yang kuat antara rasio murid per guru dengan Angka Partisipasi Murni (APM). Dengan kata lain, wilayah dengan rasio guru yang padat ataupun longgar sama-sama memiliki variasi angka partisipasi sekolah yang beragam (ada yang tinggi, ada yang rendah).

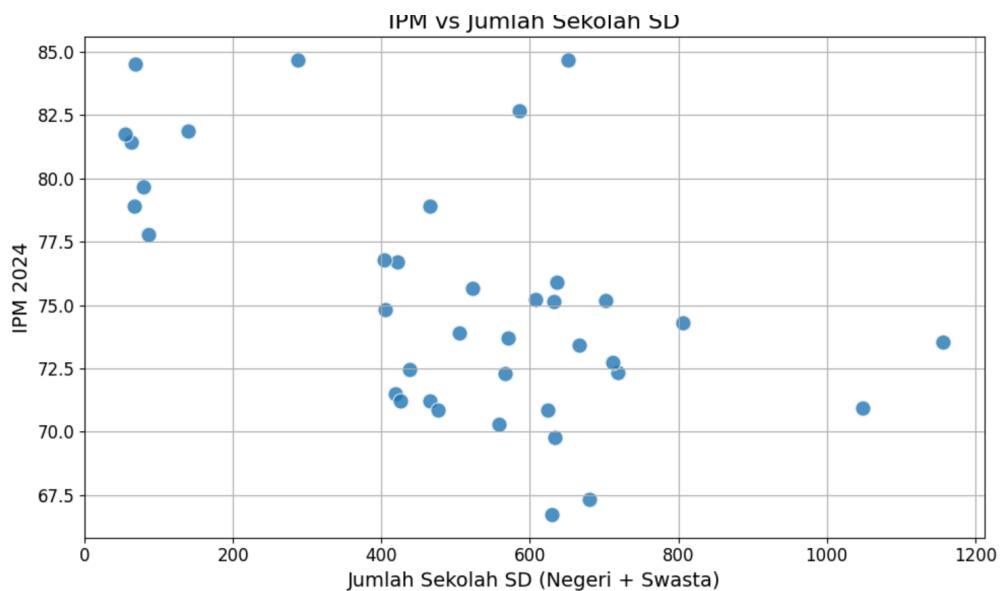
9. Hubungan APM SD vs Jumlah Sekolah di kabupaten jawa timur



Grafik ini menganalisis hubungan antara ketersediaan fasilitas (jumlah sekolah) dengan tingkat partisipasi sekolah (APM).

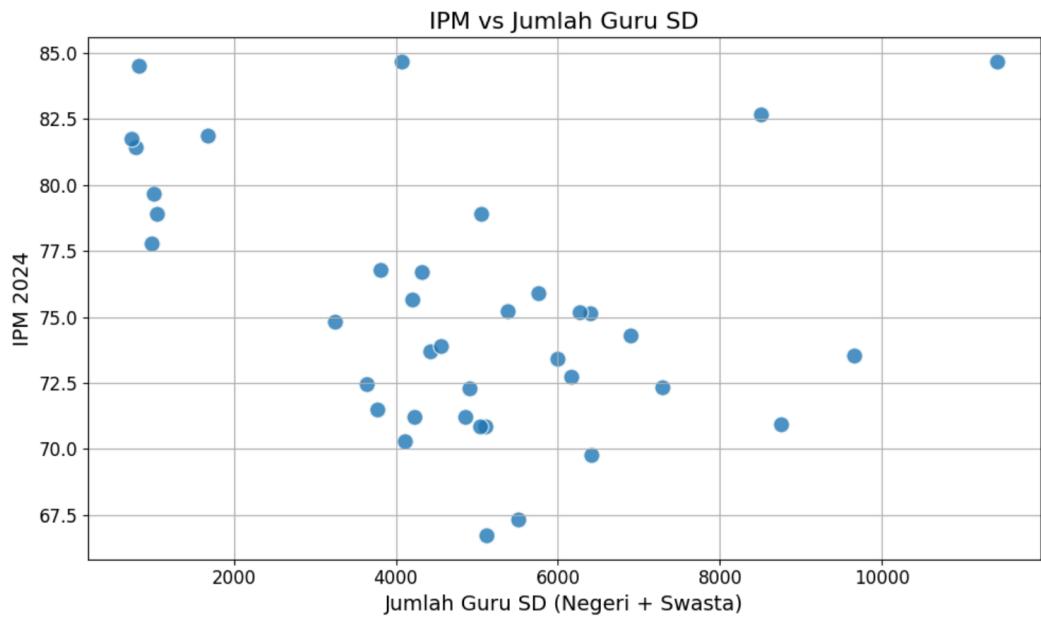
- Sumbu X: Jumlah Sekolah SD (Negeri + Swasta), dengan rentang dari 0 hingga 1.200 sekolah.
- Sumbu Y: APM SD 2024.
- Analisis:
 - Mayoritas wilayah memiliki jumlah sekolah di bawah 800 unit. Di rentang ini, nilai APM sangat bervariasi (naik-turun drastis).
 - Terdapat dua titik data (outlier) di ujung kanan (wilayah dengan >1.000 sekolah) yang justru memiliki APM di kisaran rendah (<96%).
 - Ini mengindikasikan bahwa sekadar memiliki jumlah sekolah yang banyak tidak menjamin tingginya angka partisipasi murni (APM).

10. IPM vs Jumlah Sekolah SD



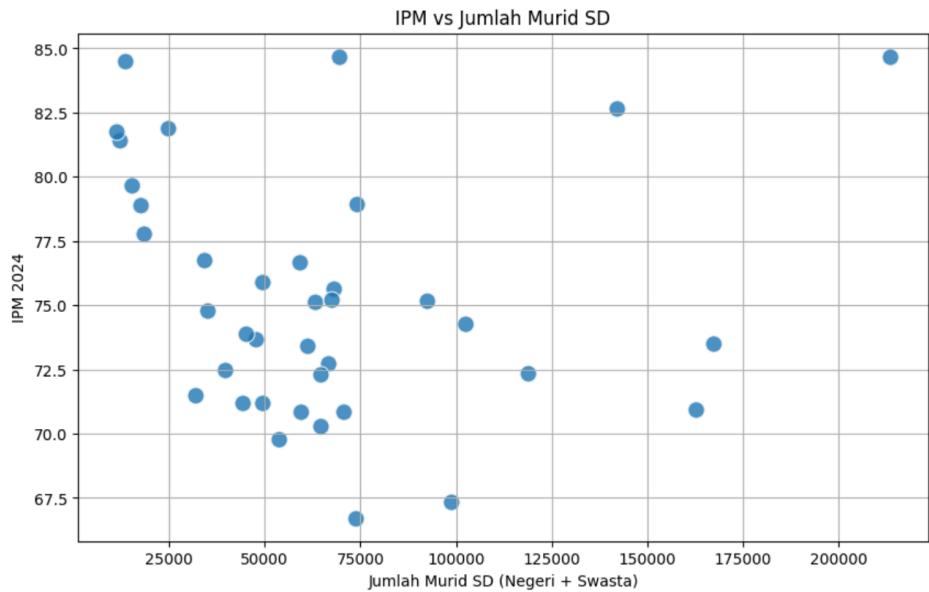
- Sumbu X: Jumlah Sekolah SD (Negeri + Swasta), rentang 0 hingga 1.200.
- Sumbu Y: Indeks Pembangunan Manusia (IPM) 2024, rentang 67.5 hingga 85.0.
- Analisis:
 - Grafik menunjukkan bahwa wilayah dengan Jumlah Sekolah SD yang sedikit (di bawah 200) memiliki IPM yang sangat tinggi (sekitar 79 hingga 85, kelompok paling atas).
 - Sebaliknya, wilayah dengan jumlah sekolah terbanyak (> 1.000) cenderung memiliki IPM yang lebih rendah (di bawah 75).
 - Ini mengindikasikan adanya korelasi negatif yang samar yang memiliki banyak sekolah (yang mungkin mengindikasikan wilayah pedesaan atau luas) dikaitkan dengan IPM yang lebih rendah.

11. IPM vs Jumlah Guru SD



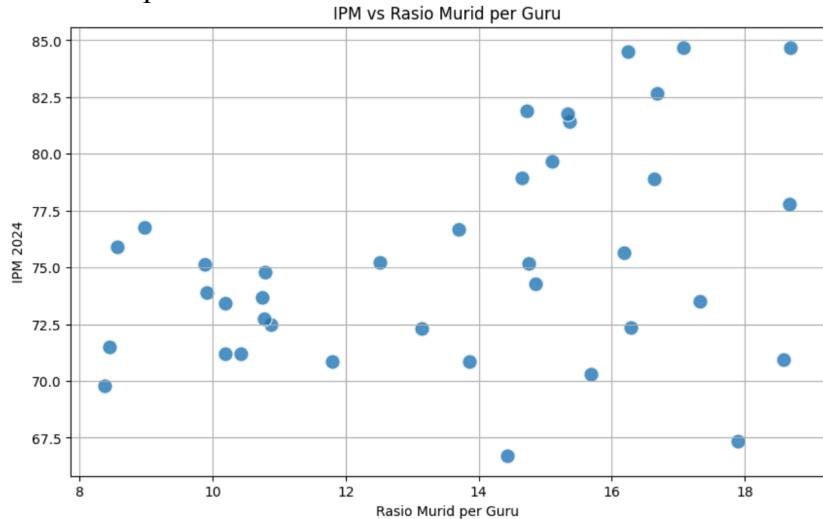
- Sumbu X: Jumlah Guru SD (0 hingga 10.000).
- Sumbu Y: IPM 2024.
- Analisis: Terdapat korelasi negatif yang samar.
 - Wilayah dengan Jumlah Guru yang sedikit (di bawah 2.000) memiliki IPM tertinggi (di atas 77.5).
 - Daerah yang memiliki jumlah guru terbanyak (9.500) memiliki IPM yang lebih rendah yaitu 73.
 - Hal ini mengindikasikan bahwa wilayah dengan sumber daya guru yang sangat banyak (biasanya di kabupaten/kota besar dan padat) tidak selalu memiliki IPM tertinggi

12. IPM vs Jumlah Murid SD



- a. Sumbu X: Jumlah Murid SD (Negeri + Swasta), rentang 25.000 hingga 200.000.
- b. Sumbu Y: IPM 2024, rentang 67,5 hingga 85.
- c. Analisis: Terdapat korelasi negatif yang samar. Wilayah dengan jumlah murid yang sedikit (di bawah 25.000) cenderung memiliki IPM yang sangat tinggi (di atas 81). Sebaliknya, wilayah dengan jumlah murid terbesar (mendekati 200.000) hanya memiliki IPM di sekitar 74. Ini menunjukkan bahwa wilayah yang padat penduduknya dan memiliki murid terbanyak tidak selalu memiliki IPM tertinggi
- d. Ada satu wilayah dengan jumlah murid sangat besar yang juga memiliki IPM sangat tinggi. Ini kemungkinan adalah kota besar atau metropolitan yang padat penduduk namun memiliki fasilitas kesehatan, ekonomi, dan pendidikan yang sangat maju.

13. IPM vs Rasio Murid per Guru



a. Hubungan Negatif

Terdapat tren penurunan rasio murid per guru seiring dengan naiknya Indeks Pembangunan Manusia (IPM). Hal ini menunjukkan bahwa wilayah dengan kualitas hidup dan pendidikan yang lebih tinggi (IPM tinggi) cenderung memiliki lebih banyak guru per murid, yang berarti perhatian pendidikan lebih personal dan sumber daya pendidikan lebih memadai.

b. Implikasi Kebijakan Pendidikan

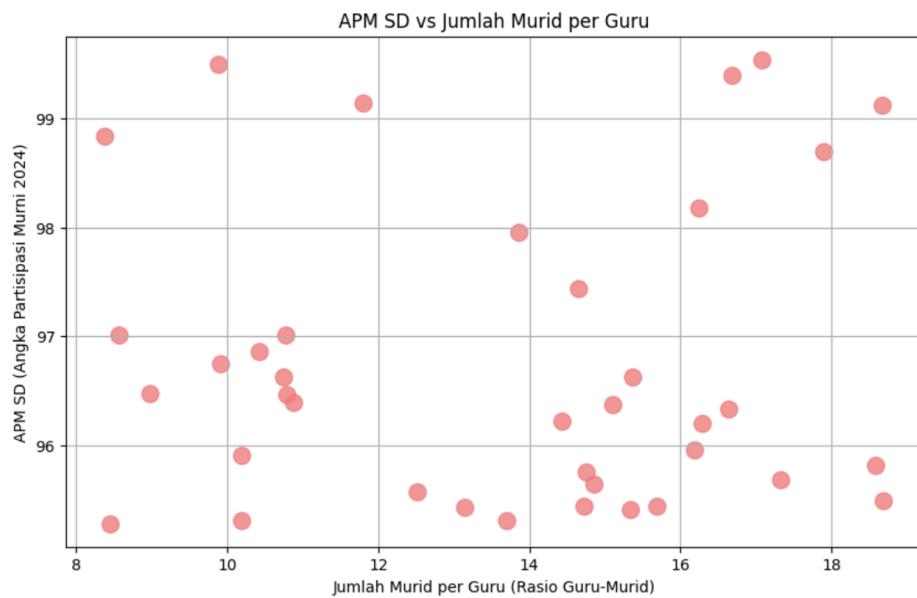
Grafik mengindikasikan bahwa mengurangi rasio murid per guru mungkin berkontribusi pada peningkatan IPM. Ini dapat menjadi dasar kebijakan untuk menambah jumlah guru atau memperbaiki distribusi guru, terutama di daerah dengan IPM rendah dan rasio murid per guru yang tinggi.

c. Ketimpangan Antar Wilayah

Sebaran titik yang cukup lebar pada rasio murid per guru untuk tingkat IPM yang serupa (misalnya di IPM 12-14) mencerminkan variasi kondisi pendidikan di berbagai daerah. Ini menegaskan perlunya pendekatan kebijakan yang

kontekstual dan tidak seragam, karena faktor selain rasio guru-murid juga mempengaruhi IPM.

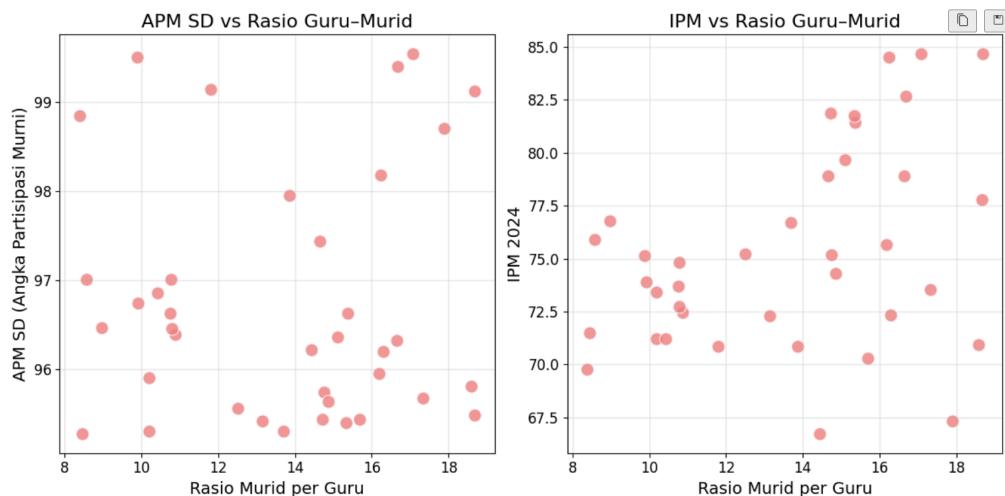
14. APM SD vs Jumlah Murid per Guru



- Titik data tersebar cukup luas, tidak membentuk pola linier yang jelas. Artinya, rasio murid per guru bukan satu-satunya faktor penentu tingginya Angka Partisipasi Murni (APM) SD. Daerah dengan rasio murid per guru serupa dapat memiliki APM yang berbeda cukup signifikan, menunjukkan adanya faktor lain seperti akses sekolah, kondisi ekonomi keluarga, dan kebijakan lokal.
- Mayoritas APM SD berada di kisaran 96–98%, termasuk di daerah dengan rasio murid per guru tinggi (misal 16–18). Ini mencerminkan bahwa program wajib belajar 9 tahun dan kebijakan inklusif pendidikan dasar di Indonesia telah berhasil menjaga partisipasi sekolah meski dengan keterbatasan jumlah guru.
- Meski sebarannya luas, terdapat kecenderungan bahwa APM tertinggi (98%) banyak ditemukan pada rasio 10–14 murid per guru.
- Rasio di atas 16 murid per guru cenderung memiliki APM sedikit lebih rendah (sekitar 96–97%). Hal ini mengisyaratkan bahwa meski APM

tetap tinggi, kualitas pembelajaran (bukan sekadar kehadiran) mungkin lebih terdampak jika rasio guru-murid terlalu besar.

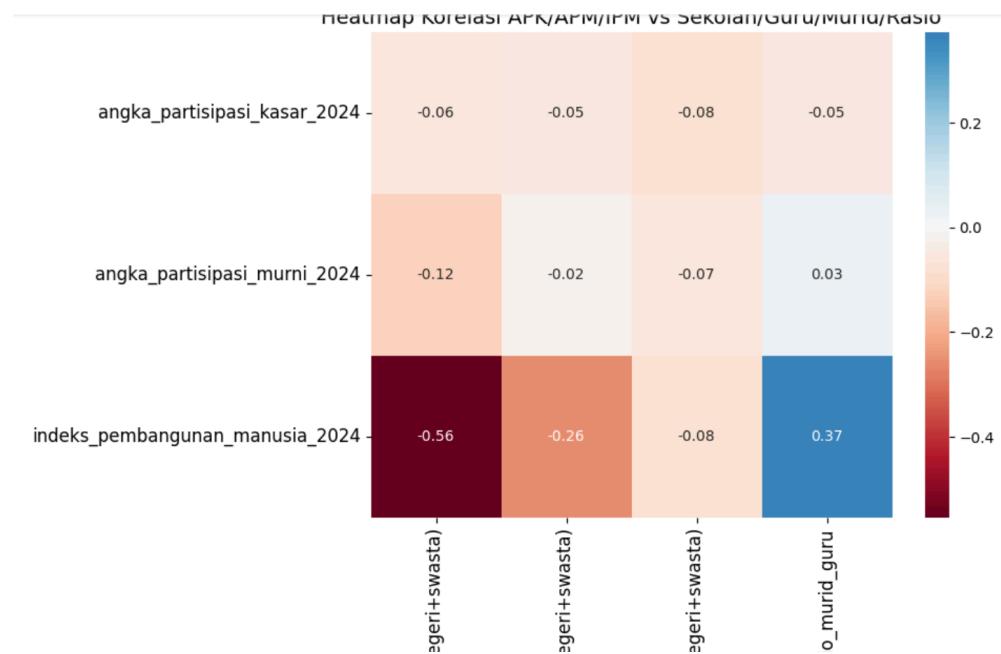
15. Perbandingan APM dan IPM vs Rasio Guru-Murid



- Rasio murid per guru berkaitan lebih erat dengan kualitas pendidikan (yang tercermin dalam IPM) daripada sekadar kehadiran di sekolah (APM SD). Oleh karena itu, kebijakan pendidikan sebaiknya tidak hanya berfokus pada meningkatkan APM, tetapi juga memperbaiki rasio guru-murid untuk mendukung pembelajaran yang lebih efektif dan dampak jangka panjang pada pembangunan manusia.
- APM SD cenderung tetap tinggi ($>96\%$) meski rasio murid per guru besar (misal 16-18). Hal ini menunjukkan bahwa akses ke sekolah dasar relatif telah tercapai, tetapi IPM yang mencakup kualitas pendidikan, kesehatan, dan daya beli lebih sensitif terhadap rasio guru-murid yang tidak ideal.
- Rasio murid per guru berkaitan lebih erat dengan kualitas pendidikan (yang tercermin dalam IPM) daripada sekadar kehadiran di sekolah (APM SD).
- Oleh karena itu, kebijakan pendidikan sebaiknya tidak hanya berfokus pada meningkatkan APM, tetapi juga memperbaiki rasio guru-murid

untuk mendukung pembelajaran yang lebih efektif dan dampak jangka panjang pada pembangunan manusia.

16. Heatmap Korelasi APK/APM/IPM vs Sekolah/Guru/Murid/Rasio



- a. IPM memiliki korelasi negatif yang signifikan dengan variabel rasio guru-murid. IPM 2024 memiliki korelasi negatif terkuat dengan jumlah murid per guru (-0.56) dan juga cukup kuat dengan rasio lainnya (-0.37). Ini menunjukkan bahwa semakin tinggi rasio murid per guru (guru lebih sedikit), semakin rendah IPM suatu daerah. Hubungan ini jauh lebih kuat dibandingkan dengan APK atau APM.

Implikasi: Kualitas pembangunan manusia (IPM) sangat dipengaruhi oleh ketersediaan guru, bukan hanya akses sekolah.

- b. APK dan APM hampir tidak berkorelasi dengan rasio guru-murid. Angka Partisipasi Kasar (APK) dan Angka Partisipasi Murni (APM) memiliki korelasi sangat lemah (mendekati nol, antara -0.02 hingga -0.12) dengan semua variabel rasio. Artinya, tingkat partisipasi sekolah (APK/APM) tidak terlalu terpengaruh oleh jumlah guru atau rasio murid per guru.

Implikasi: Kebijakan untuk meningkatkan partisipasi sekolah (kehadiran) bisa berjalan tanpa harus menambah guru, tetapi untuk meningkatkan mutu (IPM), rasio guru-murid harus diperbaiki.

- c. Korelasi terkuat hanya terjadi pada IPM, bukan pada partisipasi pendidikan Dari semua indikator (APK, APM, IPM), hanya IPM yang menunjukkan korelasi berarti dengan variabel rasio. Ini mengindikasikan bahwa dampak nyata dari rasio guru-murid terlihat pada hasil pembangunan manusia secara holistik (pendidikan, kesehatan, pendapatan), bukan sekadar pada kehadiran di sekolah.

Implikasi: Investasi dalam penambahan guru dan perbaikan rasio murid-guru akan lebih berdampak pada peningkatan IPM daripada sekadar mengejar peningkatan APK/APM

4.3 Rencana Tindak Lanjut

Setelah data bersih dan terintegrasi, beberapa langkah analisis lanjutan yang dapat dilakukan antara lain:

- a. Analisis Kualitas Pendidikan: Menyelidiki faktor-faktor yang mempengaruhi IPM, seperti rasio guru-murid, jumlah sekolah, dan indikator sosial-ekonomi lain, untuk memberikan rekomendasi kebijakan berbasis data.
- b. Analisis Klaster (Clustering): Mengelompokkan 38 kabupaten/kota di Jawa Timur berdasarkan karakteristik pendidikan menggunakan metode clustering (misal: K-Means) untuk mengidentifikasi pola kesenjangan pendidikan.
- c. Visualisasi Lanjutan: Membuat visualisasi interaktif untuk memudahkan interpretasi data dan mendukung pembuatan kebijakan.

BAB 5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan proses data wrangling, integrasi data, dan analisis eksplorasi data (EDA) terhadap data pendidikan dasar di Provinsi Jawa Timur, diperoleh beberapa kesimpulan sebagai berikut:

1. Kondisi Partisipasi Pendidikan Tinggi dan Merata

Angka Partisipasi Sekolah (APS) Provinsi Jawa Timur pada jenjang SD berada pada nilai 99,34%, menunjukkan bahwa hampir seluruh anak usia sekolah dasar telah mengakses pendidikan. Selain itu, sebagian besar wilayah memiliki APK di atas 100%, sehingga aksesibilitas pendidikan dasar tergolong sangat baik dan merata.

2. Terdapat Kesenjangan dalam Indeks Pembangunan Manusia (IPM)

Meskipun akses pendidikan merata, distribusi IPM antar kabupaten/kota menunjukkan kesenjangan. Wilayah urban seperti Kota Surabaya, Kota Malang, dan Kota Madiun mendominasi posisi IPM tertinggi, sedangkan wilayah rural dan kepulauan cenderung memiliki nilai lebih rendah. Hal ini menandakan bahwa keterjangkauan pendidikan tidak selalu paralel dengan kualitas pembangunan manusia.

3. Hubungan Antara Jumlah Sekolah dan Indikator Pendidikan Tidak Selalu Linear

Scatter plot yang membandingkan jumlah sekolah dengan indikator seperti APK, APM, dan IPM menunjukkan bahwa bertambahnya jumlah sekolah tidak otomatis meningkatkan kualitas atau partisipasi pendidikan. Faktor geografis, distribusi penduduk, dan kualitas fasilitas lebih berperan dalam mempengaruhi capaian pendidikan daerah.

4. Variasi Rasio Murid-Guru Mengindikasikan Ketimpangan Tenaga Pengajar

Analisis rasio murid per guru menunjukkan variasi signifikan antar kabupaten/kota. Beberapa wilayah memiliki rasio tinggi yang berpotensi menurunkan kualitas pembelajaran karena beban kerja guru yang berat. Hal ini menjadi indikator penting perlunya pemerataan tenaga pendidik.

5. Terdapat Korelasi yang Cukup Relevan antar Variabel Pendidikan

Heatmap korelasi memperlihatkan hubungan kuat antara variabel seperti APK, APM, jumlah murid, dan jumlah guru. Namun, hubungan antara IPM dan variabel infrastruktur (jumlah sekolah) cenderung lemah hingga negatif, menegaskan bahwa kualitas pembangunan manusia lebih dipengaruhi faktor kompleks seperti ekonomi, kualitas pengajaran, dan lingkungan sosial.

Secara keseluruhan, proyek ini menunjukkan bahwa meskipun akses pendidikan dasar di Jawa Timur sangat baik, kualitas pendidikan dan ketimpangan pembangunan manusia masih menjadi tantangan yang perlu diperhatikan dalam kebijakan pendidikan.

5.2 Saran

Berdasarkan temuan analisis pada penelitian ini, beberapa saran yang dapat diberikan adalah:

1. Saran Kebijakan Pendidikan

a. Pemerataan Distribusi Guru

Pemerintah daerah perlu memprioritaskan redistribusi tenaga pengajar ke wilayah dengan rasio murid-guru tinggi agar beban kerja guru lebih

b. proporsional dan kualitas pembelajaran lebih terjaga.

Peningkatan Kualitas Sekolah di Wilayah Rural

Meskipun jumlah sekolah banyak, beberapa wilayah IPM rendah membutuhkan peningkatan kualitas sarana pendidikan, seperti fasilitas kelas, perpustakaan, dan akses teknologi.

c. Intervensi Terarah pada Wilayah dengan IPM Rendah

Kebijakan khusus seperti bantuan sosial pendidikan, peningkatan transportasi sekolah, dan pelatihan guru dapat membantu memperbaiki capaian pembangunan manusia di daerah yang tertinggal.

2. Saran untuk Penelitian Selanjutnya

a. Menggunakan Data yang Lebih Luas dan Beragam

Penelitian berikutnya dapat menggabungkan data kualitas guru, akreditasi sekolah, atau faktor ekonomi sebagai variabel untuk melihat hubungan yang lebih komprehensif dengan IPM.

b. Menggunakan Teknik Analitik yang Lebih Lanjut

Model statistik atau machine learning (misalnya regresi linier, clustering, atau random forest) dapat digunakan untuk memberikan prediksi dan insight yang lebih mendalam terkait faktor yang paling mempengaruhi IPM atau rasio murid-guru.

c. Integrasi Data Spasial

Analisis geospasial (GIS) dapat ditambahkan untuk memetakan ketimpangan pendidikan secara visual berdasarkan letak geografis, sehingga kebijakan pemerataan dapat lebih tepat sasaran.

3. Saran untuk Implementasi Teknologi

a. Pengembangan Dashboard Pemantauan Pendidikan

Data yang telah dibersihkan dapat digunakan untuk membuat dashboard interaktif agar pemerintah dan publik dapat memantau kualitas pendidikan secara real time.

b. Automasi Data Wrangling

Tahapan pengolahan data dapat diotomatisasi menggunakan pipeline sehingga proses pembaruan data tahunan menjadi lebih cepat dan efisien.

BAB 6 KONTRIBUSI ANGGOTA KELOMPOK

1. Atya Azharina Aziz (24031554194)

- Merumuskan ide proyek.
- Mengumpulkan 3 sumber data
- Menyusun latar belakang, rumusan masalah, dan tujuan analisis.
- Mengumpulkan data dari berbagai sumber (Excel, CSV, PDF/Web).
- Membersihkan, mengintegrasikan, dan mentransformasi data.
- Melakukan analisis eksplorasi data (EDA)
- Membuat semua visualisasi: scatter plot, histogram, boxplot, bar chart.
- Menyusun dan merapikan laporan akhir.

2. Muhamad Fadly Saifullah (24031554173)

- Mengumpulkan 1 sumber data
- Melakukan web scraping.
- Merumuskan ide proyek.
- Menyusun dan merapikan laporan akhir.
- Menyusun PPT.

Daftar Pustaka

Badan Pusat Statistik Jawa Timur. (2024). *Jumlah sekolah, guru, dan murid Sekolah Dasar di bawah Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi*. Surabaya: BPS Jawa Timur.

Kementerian Sekretariat Negara Republik Indonesia. (2023). *Asta Cita: Visi dan misi pembangunan pendidikan nasional*. Jakarta: Sekretariat Negara.

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). *Wrangler: Interactive visual specification of data transformation scripts*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3363–3372). ACM.

Wickham, H. (2014). *Tidy data*. *Journal of Statistical Software*, 59(10), 1–23.
<https://doi.org/10.18637/jss.v059.i10>

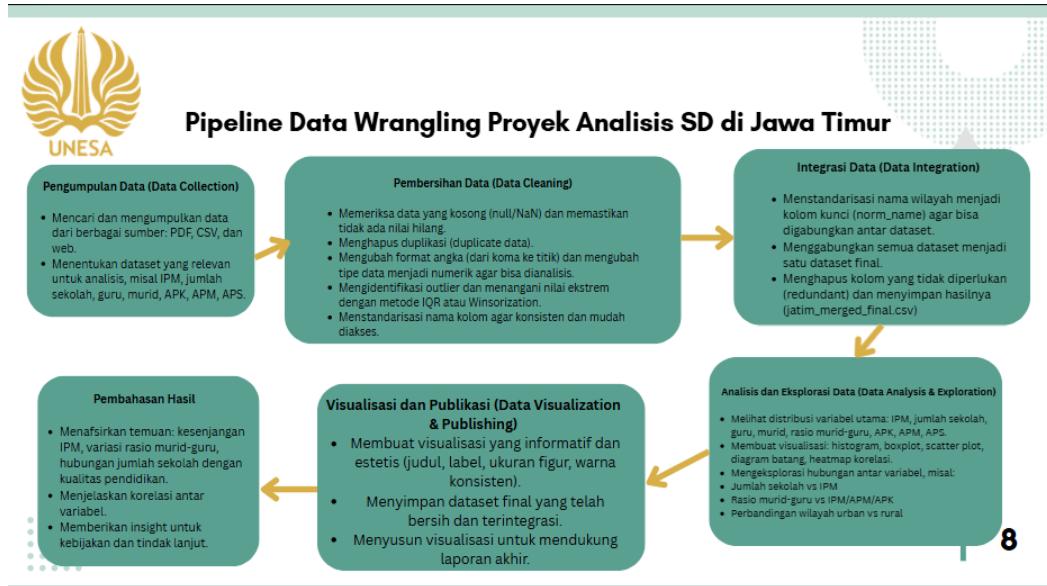
Perserikatan Bangsa-Bangsa. (2022). *Sustainable development goals report*. United Nations Publications.

Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan.

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 23 Tahun 2013 tentang Standar Penyelenggaraan Pendidikan.

LAMPIRAN

1. Pipeline Data Wrangling proyek Analisis SD di Jawa Timur



2. Repository GitHub Proyek Data Wrangling SD Jawa Timur

- Nama Repository: data-wrangling-sekolah-dasar-jatim
- Pemilik: ariez-cyber
- Link

GitHub:

<https://github.com/ariez-cyber/data-wrangling-sekolah-dasar-jatim>

● Deskripsi Repository:

Analisis data sekolah dasar di Jawa Timur untuk evaluasi rasio guru-murid dan sebaran sekolah menggunakan Python, dengan visualisasi hasil dan data wrangling terstruktur.

● Isi Repository:

- Dataset mentah
- Notebook Python untuk preprocessing, wrangling, dan analisis (DWPDF.ipynb, aps.ipynb, wrangle_jatim.ipynb, dll.)
- Laporan PDF hasil analisis dan visualisasi
- README.md sebagai panduan penggunaan

3.