PROPOSAL IDE PROJECT

MATA KULIAH DATA WRANGLING

"Tren Kualitas Udara Dunia Dan Kaitannya dengan Pembangunan Ekonomi Serta Pertumbuhan Penduduk pada Tahun 2010-2019"

Disusun oleh:

1. Nagatan Alief Putra Silahen (24031554086)
2. Khansa Nadhifa (1314622032)

Dosen Pengampun:

Ulfa Siti Nuraini, S.Stat, M.Stat.

Dr. Dian Handayani, M.Si.

Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya

S1 Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

2025

Latar Belakang Penelitian

Kualitas udara adalah isu lingkungan yang semakin mendapatkan perhatian serius karena memiliki dampak langsung terhadap kesehatan manusia yang apabila ditelusuri lebih lanjut memiliki dampak terhadap ekonomi. Partikulat halus seperti PM2.5 (partikel dengan diameter ≤ 2,5 mikrometer) merupakan polutan paling berbahaya karena dapat menembus jauh ke dalam paru-paru dan aliran darah, menyebabkan berbagai penyakit pernapasan dan kardiovaskular seperti asma, penyakit jantung, stroke, dan bahkan kematian dini.

Beban kesehatan akibat polusi udara sangat lah besar, selain itu dampak ekonomi dari polusi udara bersifat signifikan. Data global menunjukkan bahwa kerugian ekonomi dari polusi udara akibat PM2.5 dapat mencapai triliunan dolar per tahun karena kombinasi efek terhadap kesehatan, produktivitas, dan layanan kesehatan. Di sisi lain, pembangunan ekonomi dan pertumbuhan penduduk merupakan dua faktor kunci yang berpotensi berkontribusi pada polusi udara. Negara-negara dengan GDP per kapita yang tinggi sering mengalami industrialisasi, urbanisasi, dan peningkatan konsumsi energi, yang semuanya bisa meningkatkan emisi polutan. Namun, ada juga argumen bahwa dengan pertumbuhan ekonomi, negara dapat memiliki kapasitas lebih untuk mengadopsi teknologi bersih dan regulasi lingkungan yang ketat, sehingga memperbaiki kualitas udara.

Beberapa penelitian telah menunjukkan hubungan antara tingkat pendapatan GPD per kapita, kepadatan populasi, dan polusi udara. Penelitian yang telah dilakukan oleh (Kim et al., 2021) menemukan bahwa peningkatan GDP per kapita secara statistik berkorelasi dengan penurunan konsentrasi PM2.5, sementara kepadatan populasi (*population density*) berkorelasi dengan peningkatan PM2.5. Penelitian yang dilakukan oleh (Srisaringkarn & Aruga, 2025) juga menemukan bahwa Gross Provincial Product (GPP) per kapita memiliki hubungan berbentuk U dengan konsentrasi PM2.5, yang mengindikasikan pola Environmental Kuznets Curve (EKC): pada awal pembangunan ekonomi polusi meningkat, lalu menurun setelah pendapatan per kapita tertentu tercapai.

Berdasarkan temuan-temuan tersebut, sangat relevan untuk mengeksplorasi keterkaitan antara kualitas udara (PM2.5), pembangunan ekonomi (GDP per kapita), dan pertumbuhan penduduk dalam skala global. Dengan analisis data wrangling yang menggabungkan data polusi udara, GDP per kapita, dan populasi dari banyak negara, berharap penelitian ini dapat mengungkap pola-pola penting, seperti apakah pembangunan ekonomi selalu berhubungan dengan polusi, atau adakah titik di mana pertumbuhan ekonomi justru diiringi dengan perbaikan kualitas udara. Hasil analisis ini juga dapat menjadi dasar rekomendasi kebijakan pembangunan berkelanjutan: bagaimana negara bisa menyeimbangkan pertumbuhan ekonomi dengan kesehatan publik dan lingkungan.

Tujuan Penelitian

Tujuan dari penelitian ini ialah sebagai berikut:

1. Menerapkan teknik wrangling pada dataset PM2.5, GDP, dan populasi menjadi satu dataset yang siap dianalisis.

2. Mengukur hubungan antara kualitas udara (PM2.5) dengan pembangunan ekonomi (GDP per kapita) untuk mengetahui apakah peningkatan ekonomi berpengaruh pada tingkat polusi udara.

Manfaat Penelitian

Manfaat dari penelitian ini ialah sebagai berikut:

1. Menghasilkan dataset terintegrasi PM2.5, GDP, dan Populasi yang bersih dan siap digunakan untuk analisis lanjutan juga menunjukkan pentingnya pengolahan data (wrangling), integrasi, dan visualisasi dalam memecahkan masalah global.
2. Memberikan gambaran bagi pemerintah atau lembaga lingkungan tentang negara atau wilayah yang memiliki tren polusi tinggi dan faktor apa yang mungkin memengaruhinya juga dasar pertimbangan dalam merancang kebijakan pembangunan berkelanjutan yang mempertimbangkan aspek ekonomi dan lingkungan.

Proses Wrangling

1. **Import Library yang Dibutuhkan**

   import pandas as pd (untuk membaca, membersihkan, dan mengolah dataset)

   import numpy as np (untuk perhitungan numerik dan transformasi)

   import matplotlib.pyplot as plt (untuk membuat grafik dasar)

   import seaborn as sns (untuk membuat grafik yang lebih bagus dan informatif)

   plt.style.use('seaborn-v0_8') (agar style grafik tampak rapi & modern)

2. **Baca Semua dataset**

   Pada tahap ini seluruh dataset dimuat ke dalam environment untuk diproses lebih lanjut. Semua file seperti PM2.5 , GDP dan Population dibaca dan dipastikan berhasil terbuka.

   a. Dataset PM 2,5

```
pm25_raw = pd.read_csv("pm25-air-pollution.csv")
```

   Fungsi diatas yaitu untuk membaca file pm25-air-pollution.csv & Menyimpannya dalam variable pm25_raw.

```
pm25_raw.info()
```

Fungsi diatas yaitu untuk mengetahui kondisi awal dataset seperti berapa jumlah baris, kolom, tipe data tiap kolom, dan cek apa ada missing value sebelum dilakukan cleaning.

```
pm25_raw
```

Fungsi diatas yaitu untuk menampilkan seluruh isi dataframe (tabel lengkap).

b. Dataset GDP

```
gdp_raw = pd.read_csv("API_NY.GDP.PCAP.CD_DS2_en_csv_v2_252771.csv", skiprows=4)
```

Fungsi diatas yaitu untuk membaca dataset GDP per kapita dari World Bank, lalu parameter skiprows=4 karna 4 baris pertama pada file World Bank bukan data, tetapi metadata (judul, catatan, dan deskripsi). Jadi baris data sebenarnya baru dimulai pada baris ke-5, sehingga harus dilewati (skip) 4 baris tersebut, setelah itu simpan dalam variable gdp_raw.

```
gdp_raw.info()
```

Fungsi diatas yaitu untuk mengetahui kondisi awal dataset seperti menampilkan struktur dataset GDP, memastikan kolom tahun terbaca, dan melihat apakah ada banyak NaN sebelum dilakukan cleaning.

```
gdp_raw.head()
```

Fungsi diatas yaitu untuk menampilkan 5 baris pertama, membantu melihat format kolom tahun (misal: 2010,2011, dst.), dan bertujuan untuk memahami pola data GDP agar bisa di cleaning & melt

c. Dataset Population

```
population_raw = pd.read_csv("population.csv")
```

Fungsi diatas yaitu untuk membaca file mentah population.csv & Menyimpannya dalam population raw

```
population_raw.info()
```

Fungsi diatas yaitu untuk mengetahui kondisi awal dataset seperti berapa jumlah baris, kolom, tipe data tiap kolom, dan cek apa ada missing value sebelum dilakukan cleaning.

```
population_raw.head()
```

Fungsi diatas yaitu untuk menampilkan 5 baris pertama, dan bertujuan untuk memahami pola data Population agar bisa di cleaning & melt

3. Data Cleaning

Pada tahap ini berfungsi untuk membersihkan dataset dari masalah seperti duplikasi, missing values, format kolom yang tidak konsisten, dan data yang tidak relevan. Tahap ini memastikan data dalam kondisi layak untuk dianalisis.

a. Cleaning PM 2.5

```
pm25 = pm25_raw.rename(columns={
   "Entity": "Country",
    "Code": "Country_Code",
    "Concentrations of fine particulate matter (PM2.5) - Residence area type: Total": "PM25"
})

# Ambil data dari 2010 sampai 2019
pm25 = pm25[(pm25["Year"] >= 2010) & (pm25["Year"] <= 2019)]

# Save data clean
pm25.to_csv("clean_pm2,5.csv", index=False)
```

```
pm25.head()
```

Dalam cleaning PM 2.5 kami melakukan ubah nama kolom agar konsisten dengan dataset lain menggunakan fungsi rename, lalu kami filter datanya dari tahun 2010 - 2019, setelah itur kami simpan file hasil cleaning ke file CSV baru, lalu terakhir tampilkan beberapa baris awal dengan fungsi head() sebagai pengecekan akhir

b. Cleaning PM 2.5

```
# Rename kolom
gdp = gdp_raw.rename(columns={
    "Country Name": "Country",
    "Country Code": "Country_Code"
})

# Hapus kolom kosong "Unnamed: 69"
gdp = gdp.drop(columns=["Unnamed: 69"], errors="ignore")
```

```
# Mengubah dari wide ke long
gdp_long = gdp.melt(
    id_vars=["Country", "Country_Code", "Indicator Name", "Indicator Code"],
    var_name="Year",
    value_name="GDP_per_capita"
)

# Ambil data dari tahun 2010 sampai 2019
gdp_long["Year"] = gdp_long["Year"].astype(int)
gdp_long = gdp_long[(gdp_long["Year"] >= 2010) & (gdp_long["Year"] <= 2019)]

# Save data clean
gdp_long.to_csv("gdp_long.csv", index=False)

gdp_long.head()
```

Pada cleaning dataset GDP, kami mengganti nama kolom agar konsistem, lalu menghapus kolom yang tidak digunakan. Karena dataset GDP berbentuk wide (kolom tahun), kami mengubahnya menjadi long menggunakan fungsi melt() agar bisa digabung dengan dataset lain berdasarkan Year. Setelah itu kami memfilter data tahun 2010–2019, lalu menyimpannya ke file CSV baru menggunakan fungsi to_csv(), dan menampilkan beberapa baris awal dengan fungsi head() sebagai pengecekan akhir

c. Cleaning Population

```
population = population_raw.rename(columns={
    "Country Name": "Country",
    "Country Code": "Country_Code",
    "Value": "Population"
})

# Ambil data dari tahun 2010 sampai 2019
population["Year"] = population["Year"].astype(int)
population = population[(population["Year"] >= 2010) & (population["Year"] <= 2019)]

# Save data clean
population.to_csv("clean_population.csv", index=False)

population
```

Pada cleaning dataset Population, kami mengganti nama kolom agar sesuai format dataset lain, lalu mengubah kolom Year menjadi tipe integer. Setelah itu kami memfilter data untuk tahun 2010–2019. Setelah itu dataset yang sudah bersih

disimpan dalam file CSV baru menggunakan fungsi to_csv(), dan ditampilkan beberapa baris awal dengan fungsi head() sebagai pengecekan akhir.

4. **Pre-Processing**

Pada tahap ini berfungsi untuk melakukan penyesuaian struktur data seperti standarisasi nama kolom, penyesuaian tipe data, normalisasi teks, serta filtering agar dataset siap digunakan untuk proses penggabungan dan analisis.

a. Preprocessing PM2.5

```
# Buang Country_Code yang kosong (biasanya region bukan negara)
pm25_pp = pm25.dropna(subset=["Country_Code"])

# Pastikan tipe data Year benar
pm25_pp["Year"] = pm25_pp["Year"].astype(int)

print("=== PM2.5 AFTER PREPROCESSING ===")
display(pm25_pp.head())
print(pm25_pp.info())

pm25_pp.to_csv("preprocessed_pm2,5.csv", index=False)
```

Pada tahap preprocessing PM2.5, kami menghapus baris yang tidak memiliki Country_Code menggunakan fungsi dropna() karena baris tersebut umumnya bukan negara (hanya regional). Lalu kami memastikan kolom Year bertipe integer agar tidak terjadi error saat merge. Setelah itu data ditampilkan kembali menggunakan fungsi head() untuk memastikan hasilnya sesuai, lalu terakhir dataset yang sudah di preprocessing disimpan dalam file CSV baru menggunakan fungsi to_csv().

b. Preprocessing GDP per Capita

```
# Hapus region (yang Country_Code panjang 3 adalah negara)
gdp_pp = gdp_long[gdp_long["Country_Code"].str.len() == 3]

# Year harus integer
gdp_pp["Year"] = gdp_pp["Year"].astype(int)

print("=== GDP AFTER PREPROCESSING ===")
display(gdp_pp.head())
print(gdp_pp.info())

gdp_pp.to_csv("preprocessed_gdp.csv", index=False)
```

Pada preprocessing GDP, kami menyaring hanya data dengan Country_Code yang memiliki panjang 3 karakter, karena kode selain itu biasanya mewakili wilayah atau

grup ekonomi, bukan negara. Kami juga memastikan kolom Year bertipe integer agar bisa digabungkan dengan dataset lain. Hasil preprocessing kemudian dicek dengan fungsi head(), lalu terakhir dataset yang sudah di preprocessing disimpan dalam file CSV baru menggunakan fungsi to_csv().

c. Preprocessing Population

```
population_pp = population[population["Country Code"].str.len() == 3]

population_pp["Year"] = population_pp["Year"].astype(int)
population_pp["Population"] = population_pp["Population"].astype(float)

print("=== POPULATION AFTER PREPROCESSING ===")
display(population_pp.head())
print(population_pp.info())

population_pp.to_csv("preprocessed_population.csv", index=False)
```

Pada preprocessing population, kami memfilter hanya negara dengan Country_Code tiga karakter untuk menghindari duplikasi wilayah. Kami juga mengubah tipe data Year menjadi integer dan memastikan kolom populasi bertipe numerik. Hasil preprocessing kemudian dicek dengan fungsi head(), lalu terakhir dataset yang sudah di preprocessing disimpan dalam file CSV baru menggunakan fungsi to_csv().

5. Integrasi Data

Pada tahap ini berfungsi untuk menggabungkan seluruh dataset yang telah di preprocessing berdasarkan key yang sama agar menghasilkan satu dataset utama yang lengkap.

a. Merge 1: PM2.5 dengan GDP per Kapita

```
df = pm25.merge(gdp_long[["Country_Code", "Year", "GDP_per_capita"]],
                on=["Country_Code", "Year"],
                how="left")
df.head()
```

Pada tahap merge pertama, kami menggabungkan dataset PM2.5 dengan dataset GDP menggunakan kolom Country_Code dan Year. Proses ini dilakukan agar setiap negara pada tahun tertentu memiliki data PM2.5 dan GDP per kapita dalam satu tabel. Penggabungan dilakukan menggunakan metode merge(..., how="left") agar data PM2.5 tetap menjadi referensi utama. Setelah merge, beberapa baris pertama ditampilkan untuk memastikan hasil gabungan sesuai.

b. Merge 2: (PM2.5 + GDP per Kapita) dengan Dataset Population

```
final_df = df.merge(population[["Country_Code", "Year", "Population"]],
                    on=["Country_Code", "Year"],
                    how="left")
final_df
```

Setelah dataset PM2.5 dan GDP digabung, kami menambahkan dataset populasi dengan cara merge kedua menggunakan kolom Country_Code dan Year yang sama. Tujuan merge ini adalah agar satu baris data merepresentasikan tiga variabel utama: kualitas udara (PM2.5), pembangunan ekonomi (GDP per kapita), dan jumlah penduduk. Merge dilakukan dengan how="left"

```
# Hapus baris yang memiliki nilai kosong (NaN) pada kolom PM25, GDP, atau Population
final_df = final_df.dropna(subset=["PM25", "GDP_per_capita", "Population"])
final_df
```

Setelah proses penggabungan selesai, kami menghapus baris yang masih memiliki nilai kosong (NaN) pada kolom penting seperti PM25, GDP_per_capita, atau Population, menggunakan fungsi dropna(). Langkah ini memastikan dataset final benar-benar lengkap sebelum masuk tahap analisis.

```
# Menyimpan File
final_df.to_csv("dataset_final_pm25_gdp_population.csv", index=False)
```

```
final_df.info()
```

Dataset gabungan lengkap kemudian disimpan ke dalam file baru menggunakan fungsi to_csv("dataset_final_pm25_gdp_population.csv"), setelah itu hasilnya dicek kembali dengan fungsi head().

6. **Feature Engineering**

```python
# a. Hapus kolom duplikat "Country Code"
final_df = final_df.drop(columns=["Country Code"], errors="ignore")

# b. Mengurutkan berdasarkan Country dan Year
final_df = final_df.sort_values(["Country", "Year"])

# c. Menghitung selisih tahunan untuk PM2.5, GDP per kapita, dan populasi berasal dari nilai tahun sebelumnya
final_df["PM25_YoY"] = final_df.groupby("Country")["PM25"].diff()
final_df["GDP_YoY"] = final_df.groupby("Country")["GDP_per_capita"].diff()
final_df["Population_YoY"] = final_df.groupby("Country")["Population"].diff()

# d. Menghitung persentase perubahan tahunan tiap variabel.
final_df["PM25_pct"] = final_df.groupby("Country")["PM25"].pct_change()
final_df["GDP_pct"] = final_df.groupby("Country")["GDP_per_capita"].pct_change()
final_df["Population_pct"] = final_df.groupby("Country")["Population"].pct_change()

# e. Mengelompokkan nilai PM2.5
final_df["Pollution_Level"] = pd.cut(
    final_df["PM25"],
    bins=[0, 15, 35, 100],
    labels=["Low", "Moderate", "High"]
)

# f. Mengubah variabel GDP & populasi ke skala log agar grafik lebih stabil dan data tidak terlalu melebar (skew)
final_df["log_GDP"] = np.log1p(final_df["GDP_per_capita"])
final_df["log_Population"] = np.log1p(final_df["Population"])

# g. SImpan hasil future engineering ke file baru
final_df.info()
final_df.to_csv("future engineering_dataset.csv", index=False)
```

a. Menghapus kolom duplikat “Country Code”

Pada tahap pertama, kami menghapus kolom Country Code yang duplikat hasil dari proses merge. Hal ini dilakukan agar dataset lebih rapi dan tidak ada kolom yang redundan menggunakan fungsi drop().

b. Mengurutkan data berdasarkan Country dan Year

Selanjutnya data diurutkan berdasarkan kolom Country dan Year menggunakan fungsi sort_values(). berfungsi untuk mengurutkan data penting agar perhitungan perubahan tahunan (YoY) untuk tiap negara berjalan dengan benar dan tidak tercampur antara tahun.

c. Menghitung selisih perubahan tahunan (Year-over-Year)

Kami menambahkan kolom baru seperti PM25_YoY, GDP_YoY, dan Population_YoY menggunakan fungsi group by().diff(). Kami juga menggunakan fungsi diff() untuk menghitung perubahan nilai setiap tahun dibandingkan tahun sebelumnya. Tujuannya adalah untuk melihat pola naik-turun PM2.5, GDP, dan populasi pada tiap negara.

d. Menghitung persentase perubahan tahunan (%)

Setelah menghitung selisih, kami membuat persentase perubahan dengan fungsi pct_change(). Kolom baru PM25_pct, GDP_pct, dan Population_pct memberikan informasi apakah suatu nilai naik atau turun secara persentase dari tahun sebelumnya. Bagian ini penting karena untuk analisis tren yang lebih detail.

e. Mengelompokkan PM2.5 ke dalam kategori polusi

Kami menambahkan kolom Pollution_Level menggunakan fungsi pd.cut() dengan kategori: Low, Moderate, High. Pengelompokan ini mempermudah analisis dan

visualisasi pada EDA karena PM2.5 menjadi lebih mudah dibaca dan dibandingkan antar negara.

f. Transformasi logaritmik untuk GDP dan Population

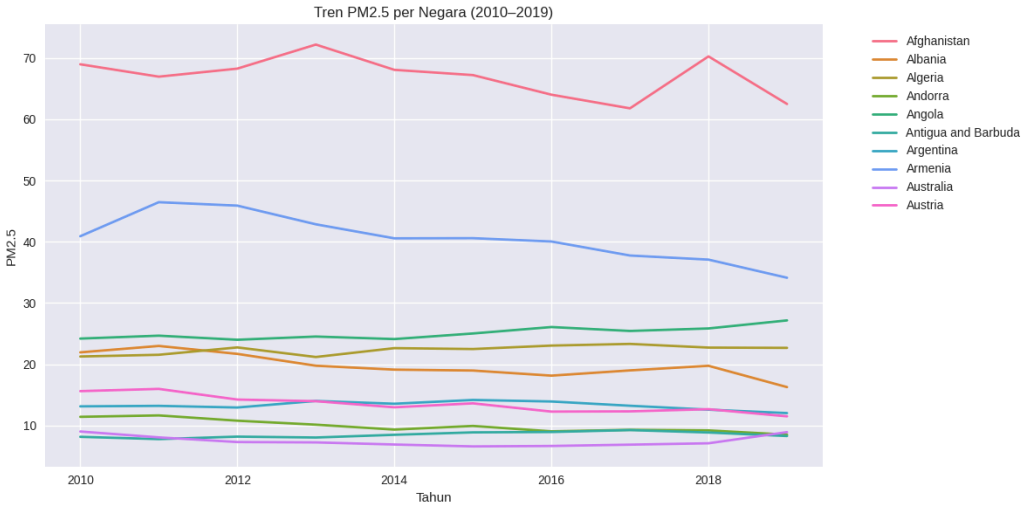
Pada bagian ini kami mengubah nilai GDP_per_capita dan Population ke dalam bentuk log menggunakan fungsi np.log1p(). Transformasi log dilakukan agar grafik lebih stabil dan tidak terlalu melebar, terutama karena nilai GDP dan Populasi antar negara bisa sangat jauh berbeda (skew).

g. Menyimpan hasil Future Engineering

Terakhir, dataset yang sudah memiliki fitur-fitur baru disimpan ke file CSV bernama future_engineering_dataset.csv menggunakan fungsi to_csv().

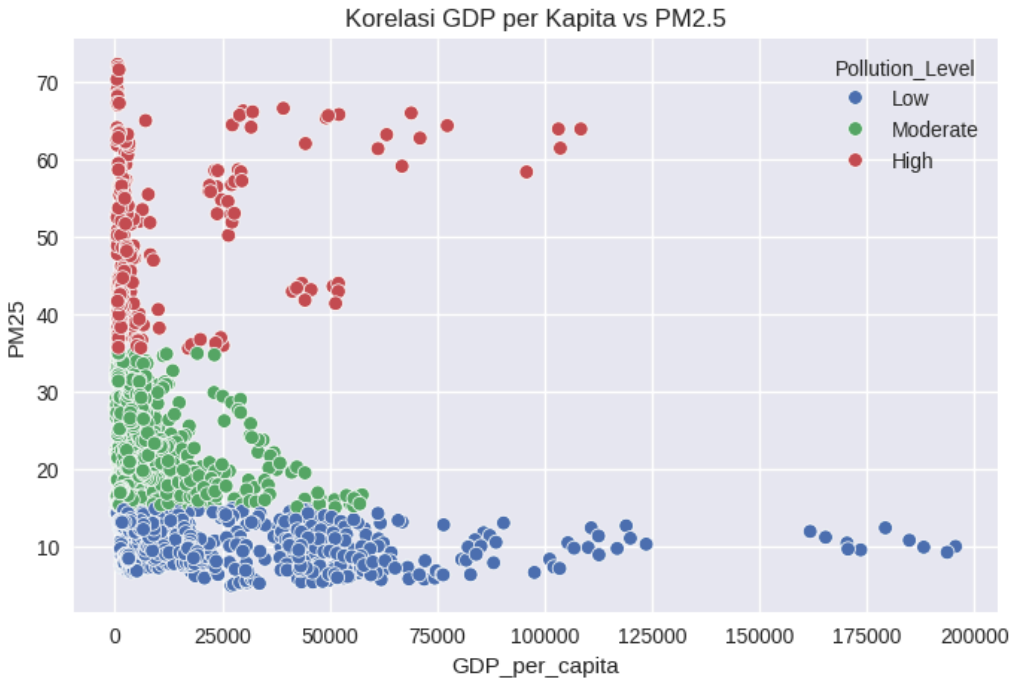
7. Exploratory Data Analysis

a. Tren Global PM2.5, GDP, dan Populasi



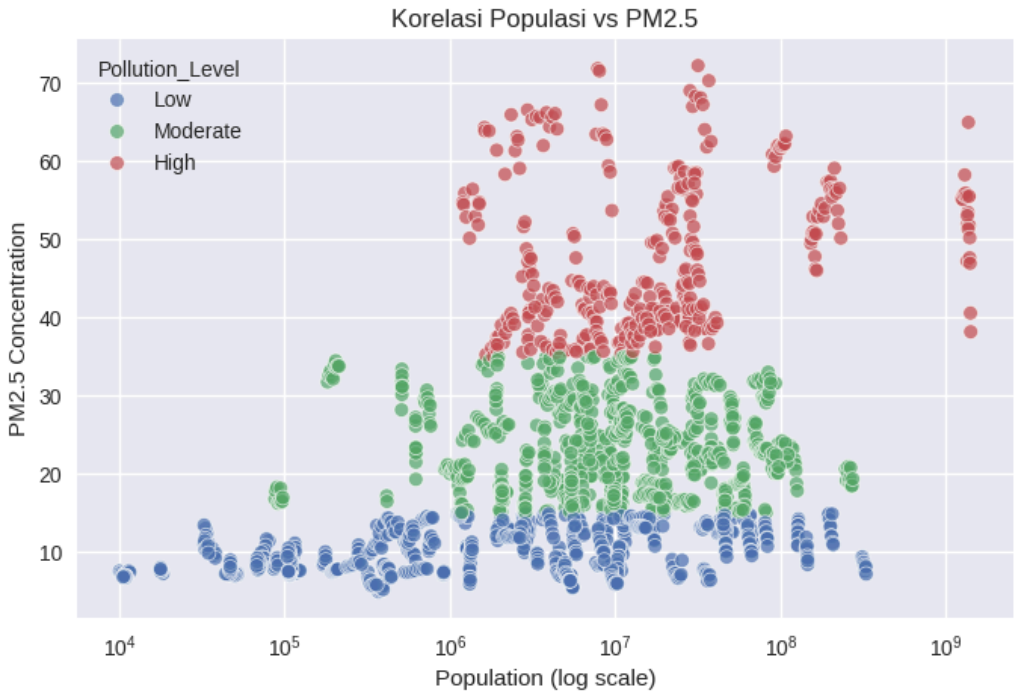
Grafik tren PM2.5 per negara dari tahun 2010 hingga 2019 diatas kami gunakan untuk melihat bagaimana perubahan kualitas udara di masing-masing negara selama satu dekade. Setiap garis mewakili satu negara, sehingga pola naik-turun PM2.5 dapat langsung terlihat. Melalui grafik ini dapat diamati bahwa beberapa negara menunjukkan penurunan polusi secara bertahap, sementara negara lain mengalami fluktuasi atau bahkan peningkatan pada tahun tertentu. Visualisasi ini membantu mengidentifikasi negara mana yang berhasil memperbaiki kualitas udaranya dan mana yang masih menghadapi tingkat polusi tinggi. Dengan demikian, grafik ini memberikan gambaran yang lebih jelas mengenai dinamika kualitas udara antar negara dan mendukung analisis perbandingan tren polusi dunia.

b. Korelasi GDP vs PM2.5



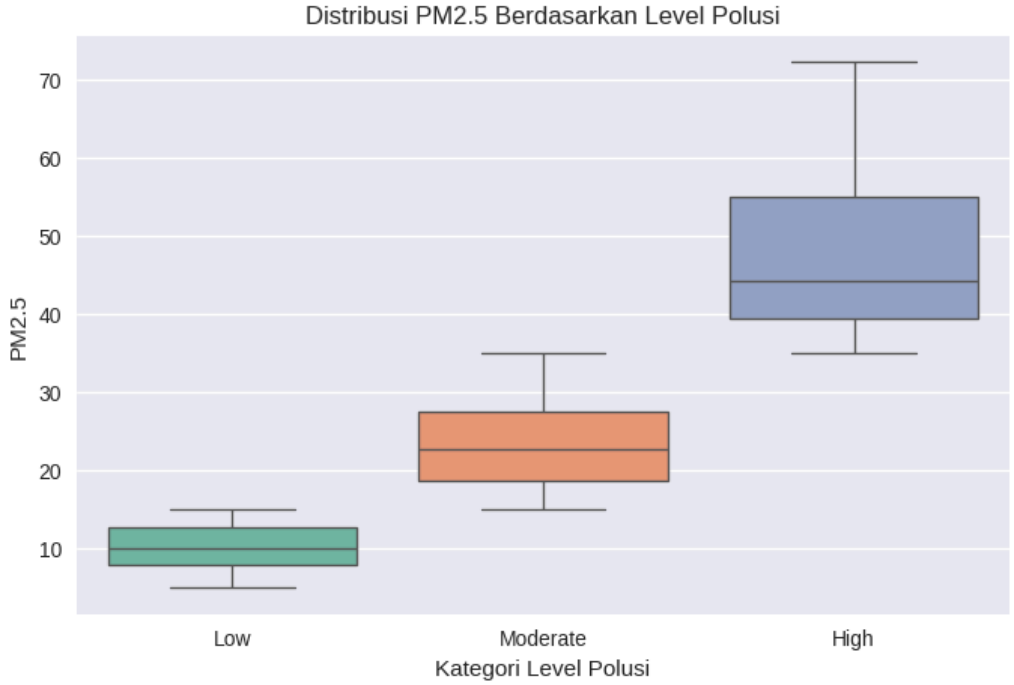
Scatterplot hubungan GDP per kapita dan PM2.5 ini digunakan untuk melihat bagaimana tingkat kemakmuran ekonomi suatu negara berhubungan dengan kualitas udaranya. Dari grafik terlihat bahwa negara dengan GDP per kapita tinggi cenderung memiliki nilai PM2.5 yang lebih rendah dan termasuk dalam kategori polusi “Low”, yang ditunjukkan oleh titik-titik berwarna biru pada bagian bawah grafik. Sebaliknya, negara yang memiliki GDP rendah lebih sering berada pada kategori polusi “High”, terlihat dari banyaknya titik merah pada area PM2.5 tinggi. Pola ini memperlihatkan hubungan negatif yang cukup jelas: semakin tinggi kemampuan ekonomi suatu negara, biasanya semakin baik pula kualitas udara yang dapat mereka capai melalui investasi teknologi bersih dan kebijakan lingkungan yang lebih efektif. Grafik ini sekaligus memperlihatkan adanya kelompok negara “Moderate” yang berada di tengah-tengah sebagai transisi antara negara berkembang dan negara maju.

c. Korelasi Population vs PM2.5



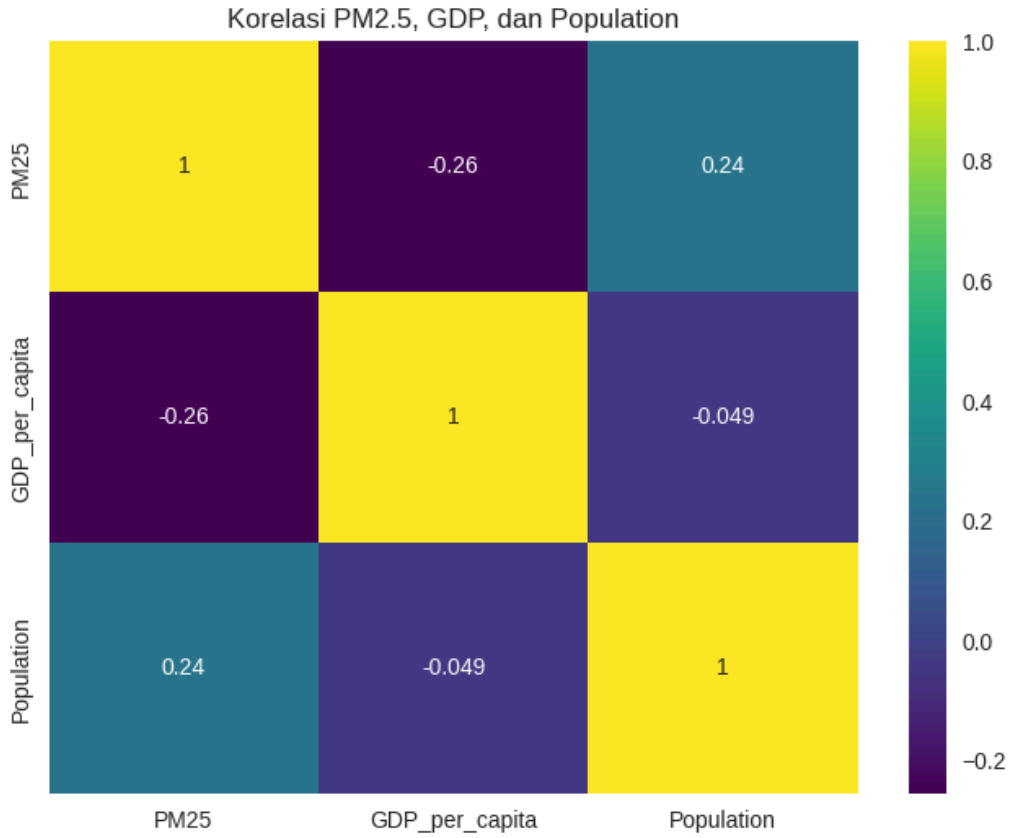
Scatterplot hubungan antara populasi & PM2.5 diatas kami gunakan untuk melihat kecenderungan apakah negara dengan jumlah penduduk yang besar memiliki tingkat polusi udara lebih tinggi. Karena populasi tiap negara berbeda jauh skala angkanya, grafik menggunakan skala logaritmik agar pola penyebaran data lebih terlihat. Visualisasi ini menunjukkan apakah pertumbuhan penduduk berkaitan dengan peningkatan PM2.5, sekaligus mengungkap pola khusus seperti negara kecil dengan polusi tinggi atau negara besar yang mampu menjaga tingkat polusi tetap rendah. Dengan demikian, grafik ini membantu memahami peran populasi terhadap kualitas udara secara lebih menyeluruh.

d. Bar plot negara dengan polusi tertinggi



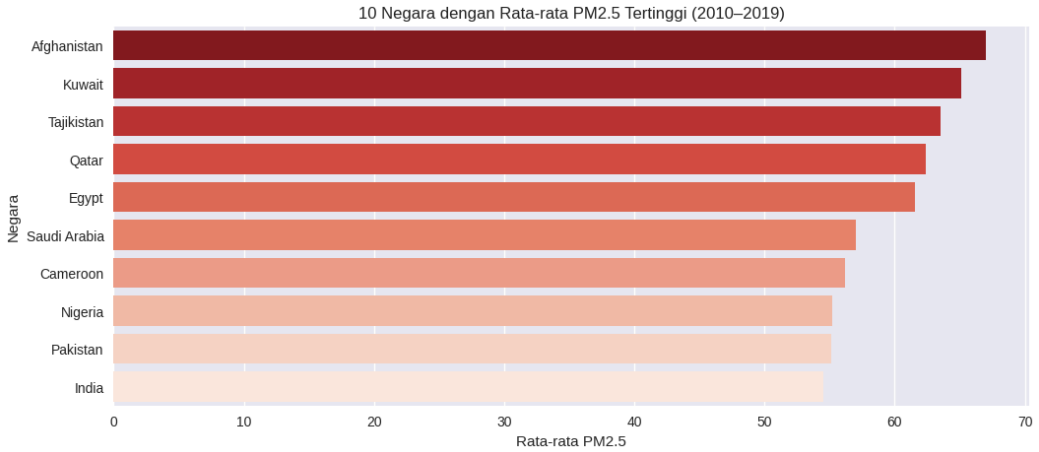
Grafik boxplot distribusi PM2.5 berdasarkan level polusi diatas kami gunakan untuk membandingkan sebaran nilai PM2.5 pada tiga kategori yaitu Low, Moderate, dan High. Grafik ini memperlihatkan bagaimana setiap kategori memiliki median, rentang nilai, dan outlier yang berbeda. Kelompok Low menunjukkan nilai PM2.5 yang rendah dan cenderung stabil, sedangkan kategori Moderate memiliki sebaran yang lebih lebar dengan nilai tengah yang lebih tinggi. Pada kategori High, terlihat bahwa nilai PM2.5 jauh lebih tinggi serta memiliki rentang yang lebih besar, menunjukkan variasi polusi yang signifikan antar negara dalam kelompok ini. Melalui visualisasi ini, perbedaan level polusi antar kategori dapat dengan jelas terlihat, sehingga mempermudah pemahaman mengenai seberapa jauh kualitas udara bervariasi antar kelompok negara.

e. Heatmap hubungan semua variabel



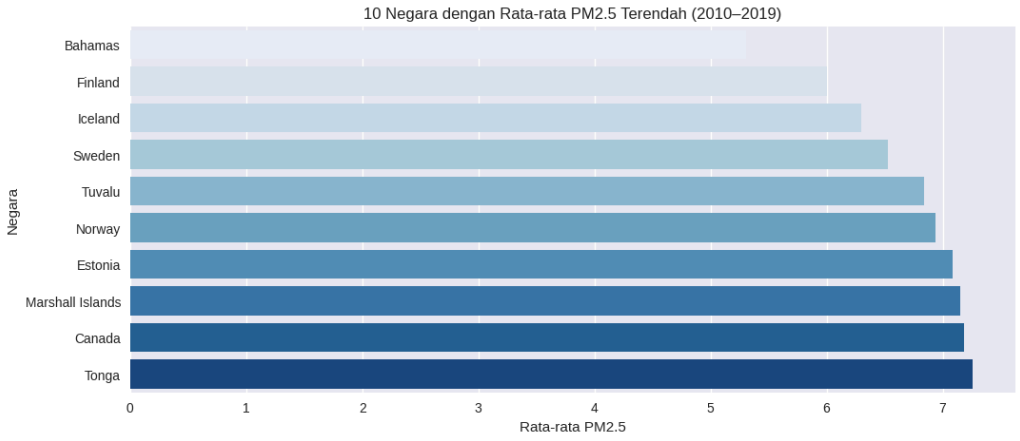
Heatmap korelasi diatas kami gunakan untuk melihat kekuatan dan arah hubungan antara tiga variabel utama, yaitu PM2.5, GDP per kapita, dan populasi. Nilai korelasi menunjukkan bahwa PM2.5 memiliki hubungan negatif dengan GDP per kapita, yang berarti semakin tinggi tingkat perekonomian suatu negara, biasanya kualitas udaranya lebih baik atau tingkat PM2.5 lebih rendah. Sebaliknya, hubungan antara PM2.5 dan populasi bernilai positif, menunjukkan bahwa negara dengan jumlah penduduk lebih besar cenderung memiliki polusi udara lebih tinggi, meskipun pengaruhnya tidak terlalu kuat. Adapun hubungan populasi dengan GDP per kapita bernilai negatif dan sangat lemah, mengindikasikan bahwa jumlah penduduk tidak berhubungan langsung dengan tingkat kemakmuran ekonomi. Melalui visualisasi ini, pola hubungan antar variabel dapat dipahami secara cepat dan membantu memperkuat interpretasi pada analisis berikutnya.

f. Plot negara dengan PM2.5 tertinggi



Grafik bar chart ini menampilkan sepuluh negara dengan rata-rata PM2.5 tertinggi selama periode 2010–2019, sehingga dapat terlihat negara mana yang mengalami tingkat polusi udara paling parah. Dari visualisasi tersebut terlihat bahwa Afghanistan memiliki nilai PM2.5 tertinggi, diikuti oleh Kuwait, Tajikistan, Qatar, dan Mesir. Negara-negara ini umumnya memiliki kondisi lingkungan yang dipengaruhi oleh faktor seperti aktivitas industri berat, urbanisasi yang cepat, penggunaan energi fosil, hingga kondisi geografis yang membuat polusi lebih mudah terperangkap. Grafik ini membantu mengidentifikasi negara-negara yang berada dalam kategori polusi ekstrem dan menjadi dasar penting untuk analisis lanjutan, seperti kaitannya dengan faktor ekonomi, populasi, atau kebijakan lingkungan yang diterapkan di masing-masing negara.

g. Plot negara dengan PM2.5 terendah



Grafik bar chart ini menampilkan sepuluh negara dengan rata-rata PM2.5 terendah selama periode 2010–2019, sehingga menunjukkan negara-negara yang memiliki kualitas udara terbaik. Negara-negara seperti Bahamas, Finland, Iceland, Sweden, dan Tuvalu berada dalam posisi teratas sebagai negara dengan tingkat polusi sangat rendah. Umumnya, negara-negara ini memiliki pengendalian lingkungan yang baik, tingkat industrialisasi yang lebih rendah, penggunaan energi bersih yang lebih dominan, serta kondisi geografis yang mendukung sirkulasi udara yang lebih sehat.

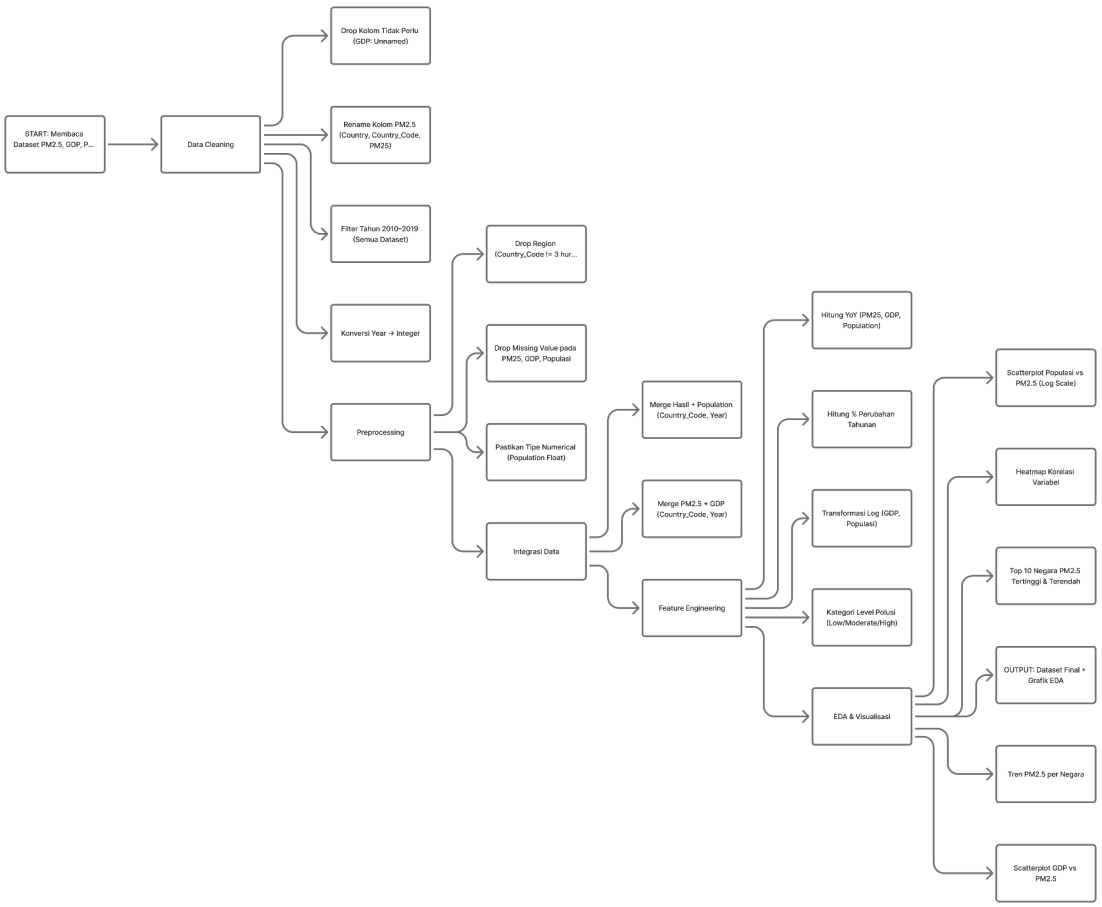
Grafik ini memberikan gambaran jelas mengenai negara-negara yang berhasil menjaga kualitas udara tetap bersih, sekaligus menjadi pembanding bagi negara-negara dengan polusi tinggi. Visualisasi ini juga memperlihatkan bagaimana perbedaan signifikan antara negara dengan sistem lingkungan yang kuat dibandingkan negara yang menghadapi tantangan polusi udara.

8. **Data Publishing**

Sebagai tahap akhir dari proses data wrangling, seluruh hasil pengolahan data, mulai dari dataset yang telah dibersihkan (cleaning), hasil pre-processing, hasil integrasi data, dataset feature engineering, hingga seluruh visualisasi EDA telah dipublikasikan (data publishing) ke dalam repository GitHub. Penerbitan data ini dilakukan agar seluruh proses dan output proyek dapat diakses, diverifikasi, serta direplikasi oleh pihak lain.

Link Github : [Nathan-090406/Projek_Data_Wrangling: Tren Kualitas Udara Dunia Dan Kaitannya dengan Pembangunan Ekonomi Serta Pertumbuhan Penduduk Tahun 2010-2019]

Diagram Alur



Dari flow pengerjaan di atas dapat dijabarkan sebagai berikut:

- Unduh dataset dari tiga sumber berbeda, yaitu dataset PM2.5, GDP per kapita, dan Populasi. Dataset diperoleh dalam format CSV, kemudian seluruh file dimasukkan ke dalam folder agar dapat dibaca oleh Google Colab/Python.
- Dataset yang berhasil dibaca oleh Google Colab kemudian dilakukan cleaning per dataset, seperti rename kolom, filtering tahun 2010–2019, dan penghapusan kolom tidak relevan. Setelah tiap dataset selesai dibersihkan, proses dilanjutkan dengan mengintegrasikan seluruh dataset berdasarkan Country_Code dan Year.
- Hasil integrasi disimpan dalam folder output dalam format CSV sebagai dataset utama yang telah tersinkronisasi.
- File CSV hasil integrasi kemudian dilakukan cleaning lanjutan dan preprocessing, yang mencakup:

  - Penghapusan data duplikat

  - Standarisasi nama kolom

  - Konversi tipe data (misalnya Year menjadi integer, Population menjadi float)

  - Handling missing value

  - Preprocessing lanjutan berupa normalisasi kolom kunci, perhitungan tren tahunan (Year-over-Year), persentase perubahan, transformasi log, dan kategorisasi level polusi (Low/Moderate/High).

- Proses cleaning dan preprocessing menghasilkan beberapa file siap olah untuk keperluan EDA (Exploratory Data Analysis). Dari dataset tersebut, dilakukan serangkaian EDA untuk memahami pola dan hubungan antarvariabel, seperti:

  - Tren PM2.5 per negara

  - Korelasi antara GDP per kapita dan PM2.5

  - Korelasi antara populasi dan PM2.5

  - Identifikasi negara dengan tingkat polusi tertinggi dan terendah

  - Analisis pola level polusi berdasarkan kategori dan distribusinya

- Tujuan utama project berhasil dicapai melalui tahap EDA, yaitu memahami hubungan PM2.5, GDP, dan Populasi serta memvisualisasikan pola dan tren global dari tahun 2010 hingga 2019.
- Langkah terakhir adalah menyimpan seluruh output, visualisasi, serta file hasil feature engineering, dan mengunggah dokumentasi project ke repository GitHub.

Saran Analisis Lanjutan

STochastic Impacts by Regression on Population, Affluence, and Technology atau STIRPAT merupakan analisis lanjutan yang disarankan untuk menganalisis pengaruh populasi dan tingkat kemakmuran ekonomi terhadap perubahan kualitas udara, khususnya konsentrasi PM2.5.

Sebelum pemodelan STIRPAT dilakukan akan diawali dengan roses integrasi dan pembersihan data dari tiga sumber utama, yaitu PM2.5, GDP per kapita, dan populasi, sehingga seluruh variabel memiliki struktur yang konsisten pada tingkat negara dan tahun. Tahapan lengkap penelitian ini adalah sebagai berikut:

1. **Pengumpulan dan integrasi Data**
   Pada tahap ini, tiga dataset utama PM2.5, GDP per kapita, dan populasi digabungkan berdasarkan kode negara dan tahun. Seluruh data dijadikan panel data periode 2010–2019, kemudian dibersihkan dari nilai hilang serta disesuaikan format datanya agar siap dianalisis.
2. **Transformasi Variabel ke Bentuk Logaritma**
   Model STIRPAT bekerja dalam bentuk logaritmik, sehingga variabel PM2.5, populasi, dan GDP per kapita ditransformasikan menjadi log_PM25, log_Population, dan log_GDP. Transformasi ini membantu menstabilkan variansi, memperbaiki distribusi data, serta memungkinkan interpretasi koefisien sebagai elastisitas.
3. **Penyusunan Model STIRPAT Dasar**
   Model dasar STIRPAT diformulasikan dalam bentuk regresi log-log:
   $log(PM2.5) = \beta_0 + \beta_1 \cdot log(Population) + \beta_2 \cdot log(GDP\ per\ capita) + \varepsilon.$
   Model ini digunakan untuk mengukur seberapa besar perubahan persentase PM2.5 dipengaruhi oleh perubahan persentase populasi dan GDP per kapita.
4. **Pengestimasian Menggunakan Regresi Panel (Fixed Effects)**
   Karena data memiliki struktur panel (negara × tahun), pemodelan dilakukan dengan regresi panel. Pendekatan Fixed Effects digunakan untuk mengendalikan perbedaan karakteristik tetap antarnegara, seperti kebijakan lingkungan, tingkat industrialisasi, dan kondisi geografis, yang dapat mempengaruhi polusi secara konstan dari waktu ke waktu.
5. **Evaluasi dan Interpretasi Hasil Model**
   Tahap ini mencakup analisis koefisien β_1 dan β_2 untuk mengetahui arah dan besarnya pengaruh populasi serta pertumbuhan ekonomi terhadap polusi udara. Koefisien dalam model log-log ditafsirkan sebagai elastisitas, sehingga memberikan gambaran mengenai perubahan proporsional PM2.5 akibat perubahan variabel lain.

Daftar Pustaka :

Jurnal :

Kim, M.-J., Chang, Y.-S., & Kim, S.-M. (2021). Impact of income, density, and population size on PM2.5 pollutions: A scaling analysis of 254 large cities in six developed countries. *International Journal of Environmental Research and Public Health, 18*(17), 9019. https://doi.org/10.3390/ijerph18179019

Silvia, S., Goembira, F., Ihsan, T., Lestari, R. A., & Irfan, M. (2020). Analisis risiko kesehatan lingkungan akibat pajanan logam dalam PM2.5 pada masyarakat di Perumahan Blok D Ulu Gadut Kota Padang. *Dampak: Jurnal Teknik Lingkungan Universitas Andalas, 17*(2), 1–10. https://doi.org/10.25077/dampak.17.2.1-10.2020

Musa, M., Rahman, P., Saha, S. K., Chen, Z., Ali, M. A. S., & Gao, Y. (2024). *Cross-sectional analysis of socioeconomic drivers of PM2.5 pollution in emerging SAARC economies*. Scientific Reports, 14, 16357. https://doi.org/10.1038/s41598-024-67199-z

Dataset :

Ritchie, H., Roser, M., & Rosado, P. (2023). *PM2.5 air pollution* [Data set]. Our World in Data. https://ourworldindata.org/grapher/pm25-air-pollution.csv

World Bank. (2024). *GDP per capita (current US$)* [Data set]. World Bank Open Data. https://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.CD?downloadformat=csv

DataHub. (2023). *Global population dataset* [Data set]. DataHub. https://datahub.io/core/population/r/population.csv