

**LAPORAN AKHIR PROJECT DATA WRANGLING
ANALISIS HUBUNGAN ANTARA LUAS LAHAN PERTANIAN,
PRODUKSI PANGAN DAN KETERSEDIAAN KALORI PER KAPITA
DI DELAPAN NEGARA ASIA TENGGARA TAHUN 2000-2021**



Dosen Pengampu:

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

Disusun oleh kelompok 5:

1. Cintiya Agustin Nareswari (24031554218)
2. Ayda Syifa Ul Aliyah (1314623064)

ABSTRAK

Penelitian ini bertujuan untuk menganalisis keterkaitan antara luas lahan pertanian, produksi pangan, dan kalori per kapita pada delapan negara Asia Tenggara dalam rentang waktu 2000–2021. Data yang digunakan berasal dari tiga sumber utama, yaitu World Bank (luas lahan pertanian), Our World in Data (kalori per kapita), dan World Food Production (produksi tujuh komoditas pangan). Seluruh dataset melalui proses pembersihan data seperti standarisasi nama negara, penyesuaian format kolom, pemilihan tahun analisis, penggabungan dataset, serta pengisian nilai hilang menggunakan interpolasi dan median. Hasil eksplorasi menunjukkan bahwa setiap negara memiliki karakteristik yang berbeda, baik dari segi luas lahan, jumlah produksi, maupun tingkat ketersediaan kalori. Indonesia dominan dalam produksi beras dan jagung, Malaysia unggul dalam minyak sawit, Kamboja memiliki produksi ubi jalar yang tinggi, sementara Malaysia dan Thailand konsisten berada pada tingkat kalori per kapita tertinggi. Analisis korelasi Spearman menunjukkan hubungan yang cenderung lemah hingga sedang antarvariabel, sehingga peningkatan pada satu variabel belum tentu diikuti oleh variabel lainnya. Selanjutnya, analisis clustering menggunakan metode K-Means menghasilkan tiga kelompok ketahanan pangan, yaitu kategori tinggi (Thailand dan Filipina), kategori sedang (Indonesia, Malaysia, Myanmar), dan kategori rendah (Vietnam, Kamboja, Timor Leste). Secara keseluruhan, penelitian ini memberikan gambaran bahwa ketahanan pangan negara-negara Asia Tenggara dipengaruhi oleh kombinasi kondisi lahan, kapasitas produksi, dan tingkat konsumsi energi masyarakat.

DAFTAR ISI

ABSTRAK.....	i
DAFTAR ISI.....	ii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	1
1.3 Tujuan	2
1.4 Manfaat	2
BAB II TINJAUAN PUSTAKA.....	3
2.1 Data Wrangling	3
2.2 Library dalam Analisis Data	3
2.3. Metode Data Wrangling.....	3
BAB III METODOLOGI PENELITIAN	7
3.1 Dataset.....	7
3.2 Flowchart Proses Data Wrangling	8
BAB IV PEMBAHASAN.....	9
4.1 Teknik Pengambilan Data	9
4.2 Data Cleaning.....	10
4.3 Integrasi Data	12
4.4 Analisis Data Eksplorasi	12
4.5 Kendala dan Rencana.....	45
4.6 Kontribusi Kelompok.....	45
BAB V PENUTUP.....	47
5.1 Kesimpulan	47
BAB IV DAFTAR PUSTAKA	48
Lampiran 1.	49

BAB I

PENDAHULUAN

1.1 Latar Belakang

Ketahanan pangan merupakan aspek penting bagi negara-negara di dunia salah satunya di Asia Tenggara. Kawasan ini memiliki pertumbuhan penduduk yang cepat serta tingkat ketergantungan yang tinggi pada sektor pertanian. Kemampuan suatu negara untuk menyediakan pangan secara berkelanjutan dipengaruhi oleh dua komponen utama yaitu luas lahan pertanian dan kapasitas produksi komoditas pangan. Keduanya berperan langsung dalam menentukan tingkat ketersediaan pangan bagi masyarakat yang salah satunya diukur melalui jumlah kalori perkapita.

Proyek ini berfokus pada delapan negara Asia Tenggara, yaitu Indonesia, Malaysia, Thailand, Vietnam, Filipina, Myanmar, Kamboja, dan Timor Leste. Negara-negara tersebut dipilih karena memiliki karakteristik pertanian yang beragam, mulai dari sektor pertanian yang sangat besar hingga negara dengan kapasitas produksi yang relatif terbatas.

Variabel produksi pangan yang digunakan pada proyek penelitian ini mencakup komoditas utama terhadap sumber pangan maupun sektor pertanian diantaranya adalah, produksi jagung, produksi beras, produksi ubi jalar, produksi minyak sawit, produksi kopi, produksi kakao, dan produksi teh. Pemilihan komoditas tersebut didasarkan pada perannya dalam sistem pangan, baik sebagai bahan pangan utama, sumber kalori tambahan, maupun komoditas perkebunan yang memiliki nilai ekonomi tinggi.

Selain itu, indikator luas lahan pertanian digunakan untuk melihat seberapa besar kapasitas struktur pertanian pada suatu negara. Data mengenai ketersediaan kalori perkapita digunakan sebagai representasi kesejahteraan pangan masyarakat yang secara langsung dipengaruhi oleh produksi pangan domestik dan ketersediaan sumber energi.

Seluruh data dikumpulkan dari sumber terpercaya seperti World Bank untuk data lahan pertanian, World Food Programme Kaggle untuk data produksi pangan, dan data nutrisi dari sumber global seperti FAO/WFP. Periode penelitian 2000–2021 dipilih untuk memungkinkan analisis tren jangka panjang serta melihat perubahan dinamika hasil produksi dan konsumsi kalori.

Melalui penggabungan data lintas negara dan lintas komoditas tersebut, project ini bertujuan untuk memahami hubungan antara luas lahan, produksi pangan, dan ketersediaan kalori, serta melihat pola karakteristik negara-negara di Asia Tenggara dalam konteks ketahanan pangan.

1.2 Rumusan Masalah

1. Bagaimana perkembangan luas lahan pertanian, produksi komoditas pangan (beras, jagung, minyak sawit, kopi, kakao, teh, ubi jalar), dan ketersediaan kalori per kapita pada delapan negara Asia Tenggara selama tahun 2000–2021?
2. Apakah terdapat hubungan antara luas lahan pertanian, produksi pangan, dan ketersediaan kalori?

3. Bagaimana perbedaan pola ketahanan pangan antar delapan negara Asia tenggara berdasarkan ketiga variabel tersebut?
4. Bagaimana pengelompokan delapan negara Asia Tenggara berdasarkan luas lahan pertanian, produksi pangan, dan ketersediaan kalori?

1.3 Tujuan

1. Mendeskripsikan tren luas lahan pertanian, produksi komoditas pangan, dan ketersediaan kalori per kapita pada delapan negara Asia Tenggara selama tahun 2000–2021
2. Menganalisis hubungan dasar antara luas lahan, produksi pangan, dan tingkat kalori menggunakan korelasi dan visualisasi.
3. Mengidentifikasi pola tren produksi dan kalori antar delapan negara Asia Tenggara.
4. Mengelompokkan delapan negara Asia Tenggara berdasarkan kesamaan karakteristik produktivitas dan kalori menggunakan metode clustering sederhana (K-Means).

1.4 Manfaat

1. Memberikan gambaran perkembangan sektor pertanian di delapan negara Asia Tenggara melalui analisis tren luas lahan, produksi komoditas utama, dan tingkat kalori per kapita selama 2000–2021.
2. Menyediakan informasi mengenai hubungan antarvariabel pertanian, seperti keterkaitan antara luas lahan, besarnya produksi komoditas, dan kontribusinya terhadap ketersediaan kalori masyarakat.
3. Mengungkap perbedaan karakteristik pertanian antarnegara, sehingga terlihat pola masing-masing negara dalam memanfaatkan lahan dan mengelola komoditas pangannya.
4. Memberikan dasar pengelompokan negara berdasarkan kesamaan luas lahan, pola produksi, dan tingkat kalori, yang dapat digunakan untuk memahami posisi relatif tiap negara dalam konteks produktivitas pertanian.
5. Menyediakan insight untuk mendukung analisis lanjutan, seperti peramalan produksi dan kalori, perencanaan sektor pertanian, serta pengembangan strategi peningkatan kapasitas produksi pangan.

BAB II TINJAUAN PUSTAKA

2.1 Data Wrangling

Data wrangling merupakan proses pengolahan data mentah menjadi data yang lebih terstruktur, rapi, dan siap digunakan untuk analisis. McKinney (2022) dalam bukunya *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* menjelaskan bahwa data wrangling melibatkan serangkaian langkah seperti pembersihan data, penggabungan dataset, transformasi format, dan penanganan nilai hilang agar data siap digunakan pada tahapan analisis lanjutan. Tahap ini sangat penting karena kualitas analisis statistik sangat bergantung pada kualitas data yang digunakan. Sejalan dengan Wickham (2014), data yang rapi (*tidy data*) harus memiliki struktur yang konsisten agar lebih mudah diproses dan dianalisis secara efektif.

2.2 Library dalam Analisis Data

Penelitian ini menggunakan beberapa library Python yang berperan penting dalam proses pengolahan dan analisis data. Pandas digunakan sebagai library utama untuk membaca, membersihkan, dan mengintegrasikan data dari berbagai sumber melalui struktur *DataFrame* yang fleksibel. Library NumPy mendukung berbagai operasi numerik seperti perhitungan log transformasi serta penjumlahan variabel produksi. Untuk visualisasi, digunakan Matplotlib dan Seaborn yang berfungsi menampilkan distribusi, tren tahunan, dan hubungan antar variabel dalam bentuk grafik informatif seperti *line plot*, *histogram*, *scatter plot*, dan *heatmap*.

Pada tahap analisis lanjutan, penelitian ini memanfaatkan Scikit-Learn untuk melakukan normalisasi data, penerapan algoritma clustering K-Means, evaluasi model dengan *Silhouette Score*, serta reduksi dimensi menggunakan PCA guna memvisualisasikan hasil clustering. Kombinasi seluruh library ini memungkinkan proses data wrangling dan analisis berlangsung secara sistematis, akurat, dan efisien. (McKinney, 2022)

2.3. Metode Data Wrangling

1. Data Cleansing

Data cleansing merupakan proses untuk mendeteksi dan memperbaiki nilai yang hilang, tidak konsisten, atau tidak akurat dalam dataset agar data memiliki kualitas yang memadai untuk dianalisis. Proses ini mencakup penanganan *missing values* menggunakan metode seperti *mean*, *median*, atau interpolasi, penghapusan duplikasi data, normalisasi format penulisan, serta penyesuaian tipe data agar seragam. Tahap ini penting karena kualitas hasil analisis sangat bergantung pada kualitas data yang digunakan (Rahm & Do, 2000).

2. Standarisasi dan Integrasi Data

Standarisasi data merupakan proses penyelarasan format, nama variabel, dan struktur penyimpanan data agar konsisten dan mudah dianalisis. Menurut Wickham (2014), data yang baik harus memiliki format yang terstandarisasi sehingga setiap variabel hanya memiliki satu arti dan satu bentuk penulisan. Setelah proses standarisasi dilakukan, langkah berikutnya adalah integrasi data, yaitu menggabungkan beberapa dataset dari berbagai sumber menjadi satu struktur yang saling terhubung. Tahapan ini penting karena memastikan seluruh informasi tersedia dalam satu bentuk yang rapi dan siap dianalisis lebih lanjut.

3. Analisis Data Eksploratif

Menurut Heckert dan Filiben (2003), analisis data eksploratif adalah sebuah pendekatan dalam menganalisis data yang menggunakan berbagai teknik untuk menggali informasi dari data, menemukan pola atau struktur yang tersembunyi, mengidentifikasi variabel penting, mendeteksi outlier, memeriksa asumsi dasar, membangun model sederhana, serta menentukan faktor terbaik. Analisis data eksploratif banyak mengandalkan teknik visualisasi. Hal ini karena manusia pada dasarnya memiliki kemampuan yang baik dalam mengenali informasi dari visualisasi grafik dibandingkan dengan informasi yang hanya berisi angka atau tulisan saja.

4. Transformasi Data

Hakimah et al. (2024) menyatakan bahwa transformasi data adalah proses mengubah nilai-nilai dalam dataset untuk mencapai tujuan tertentu, seperti memperbaiki bentuk distribusi. Distribusi pada data menjadi hal yang penting karena jika data tidak berdistribusi normal seringkali menyulitkan dalam analisis statistik yang mengharuskan asumsi distribusi normal. Selain itu, transformasi data dilakukan untuk mengurangi skewness atau kecondongan distribusi pada satu sisi. Dengan menggunakan transformasi, skewness dapat diminimalkan. Kemudian, transformasi data juga diperlukan jika suatu analisis mengharuskan memenuhi asumsi-asumsi tertentu. Biasanya uji-uji parametrik memerlukan asumsi distribusi normal. Dengan transformasi data dapat membantu meningkatkan keandalan hasil analisis.

Jenis teknik transformasi data yang dapat dilakukan adalah transformasi log. Transformasi log atau transformasi logaritma efektif dalam mengatasi data yang memiliki skewness atau sebaran data tidak merata. Transformasi logaritma dapat meredakan variabilitas data dan menjadikannya mendekati distribusi normal.

5. Z-Score Standardization

Menurut Gravetter dan Wallnau (2017) *Z-Score Standardization* adalah teknik standarisasi data yang mengubah setiap nilai asli dalam sebuah distribusi menjadi dalam bentuk nilai baku z-score. Standarisasi ini menggunakan rumus $z = (X - \mu) / \sigma$. *Z-Score Standardization* adalah metode standarisasi yang paling umum dilakukan dan

memiliki beberapa karakteristik. Pertama, *Z-Score Standardization* tidak mengubah bentuk distribusi. Hal ini terjadi karena proses standarisasi tidak mengubah posisi relatif setiap nilai, tetapi hanya mengubah label skala pengukurannya. Kedua, distribusi yang telah distandarisasi dengan z-score selalu memiliki mean = 0 dan standar deviasi = 1. Standarisasi sangat penting jika ingin melakukan analisis yang berbasis jarak antar data. Hal ini karena skala variabel sangat berpengaruh dalam proses analisisnya.

6. Analisis Korelasi

Menurut Gravetter dan Wallnau (2017) analisis korelasi merupakan metode statistik yang digunakan untuk menentukan arah dan mengukur tingkat hubungan linier antara dua variabel numerik. Nilai korelasi berkisar antara -1 hingga $+1$, di mana nilai yang mendekati $+1$ menunjukkan hubungan positif yang kuat dan nilai mendekati -1 menunjukkan hubungan negatif yang kuat. Dalam penelitian yang melibatkan data yang tidak berdistribusi normal, memiliki pola hubungan non-linier monotonik, atau mengandung outlier, teknik korelasi yang lebih tepat digunakan adalah korelasi Spearman. Korelasi Spearman mengukur keterkaitan berdasarkan peringkat (rank) dari nilai-nilai kedua variabel, bukan nilai mentahnya. Dengan demikian, teknik ini mampu menangkap pola hubungan naik-turun yang konsisten, meskipun tidak berbentuk garis lurus. Analisis korelasi Spearman digunakan untuk melihat pola keterkaitan awal antar variabel sebelum dilakukan analisis lebih lanjut, sehingga membantu peneliti mengetahui apakah suatu variabel memiliki kecenderungan meningkat atau menurun seiring perubahan variabel lain.

7. Clustering

Clustering merupakan salah satu metode analisis data dalam kategori *unsupervised learning* yang digunakan untuk mengelompokkan objek berdasarkan kemiripan karakteristiknya. Menurut Pratama (2019), objek yang berada dalam satu cluster memiliki tingkat kesamaan yang lebih tinggi dibandingkan objek pada cluster lain, sehingga metode ini efektif dalam mengidentifikasi pola tersembunyi dan struktur alami dalam data. Teknik clustering banyak digunakan dalam segmentasi data, analisis perilaku, serta eksplorasi data dalam berbagai bidang. Dengan adanya proses pengelompokan ini, informasi kompleks dapat disederhanakan sehingga analisis lanjutan dapat dilakukan secara lebih sistematis dan terarah.

8. K-means

K-Means merupakan salah satu algoritma clustering yang paling populer dalam analisis data, terutama karena kesederhanaan dan kecepatannya dalam memproses dataset berukuran besar. Menurut Jain (2010), K-Means bekerja dengan membagi data ke dalam sejumlah k kelompok berdasarkan jarak antara titik data dan pusat cluster (*centroid*). Proses ini dilakukan secara iteratif melalui dua tahap utama, yaitu pengelompokan data berdasarkan jarak terdekat dan pembaruan posisi centroid hingga hasilnya stabil. Dalam praktiknya, penentuan jumlah cluster dapat dibantu

menggunakan metode seperti Elbow Method atau Silhouette Score sehingga struktur pengelompokan yang dihasilkan lebih representatif.

9. Penyimpanan Data

Penyimpanan data (exporting) merupakan tahap akhir dalam proses data wrangling, di mana dataset yang telah dibersihkan dan ditransformasi disimpan ke dalam format tertentu agar dapat digunakan kembali untuk analisis lanjutan atau pemodelan. Menurut McKinney (2022), menyimpan data ke dalam file seperti .csv, .xlsx, atau format lainnya merupakan praktik penting dalam alur kerja data science untuk menjaga reusabilitas, efisiensi, dan keterlacakan proses pengolahan data.

BAB III

METODOLOGI PENELITIAN

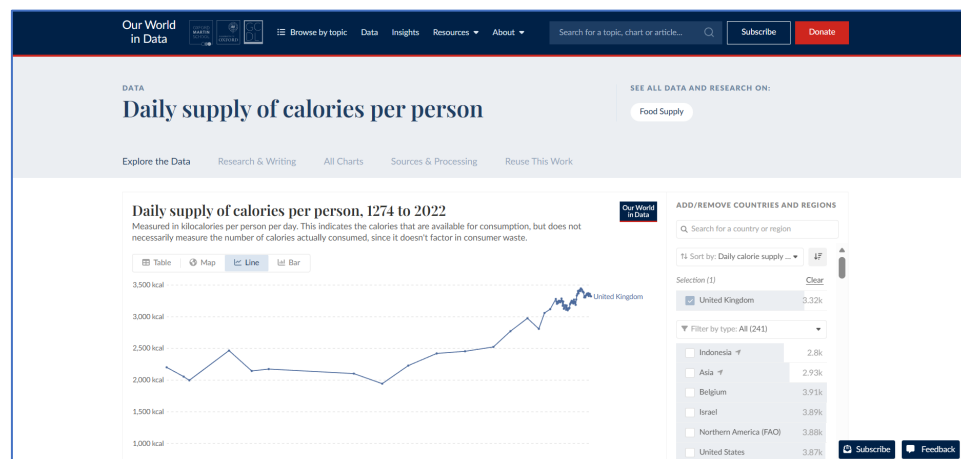
3.1 Dataset

1. World Food Production – Kaggle



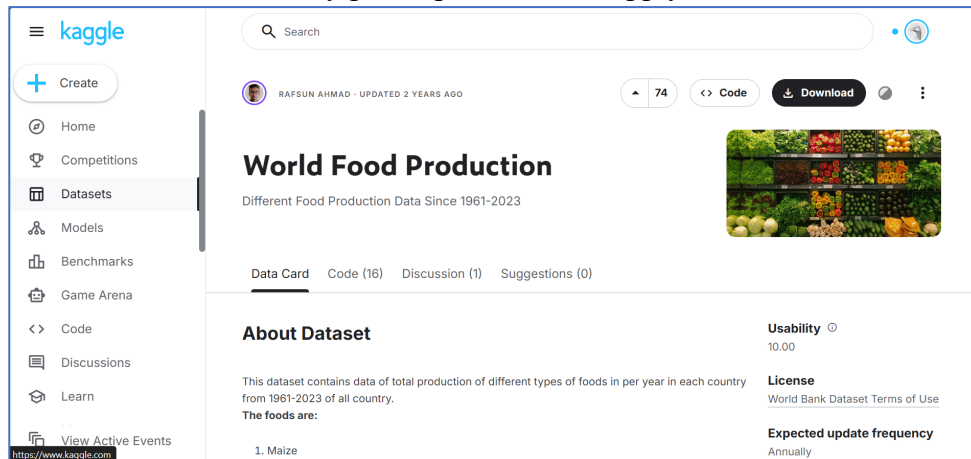
Dataset ini menyediakan produksi pangan berbagai komoditas (seperti jagung, beras, minyak sawit, kopi, kakao, teh, dan ubi jalar) dalam satuan ton untuk berbagai negara dan tahun. Data ini digunakan untuk menghitung variabel total produksi pangan dan analisis per komoditas, sehingga memungkinkan pemahaman bagaimana kapasitas produksi negara berkontribusi pada ketahanan pangan

2. World Bank – Agricultural Land (% of Land Area)



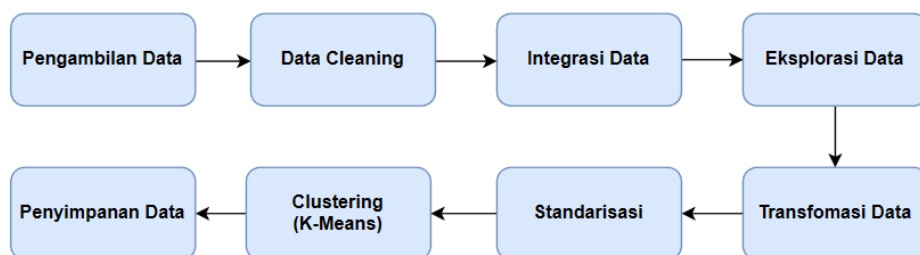
Sumber ini menyediakan persentase luas lahan pertanian dibanding total luas daratan negara (indikator AG.LND.AGRI.ZS). Data lahan pertanian penting karena merefleksikan potensi fisik produksi pangan negara serta besaran infrastruktur agrikultur yang bisa mendukung pertanian.

3. Our World in Data – Daily per Capita Caloric Supply



Dataset ini menyajikan rata-rata jumlah kalori yang tersedia per orang per hari (kalori per kapita) dalam suatu negara, yang mencerminkan tingkat ketersediaan pangan dan gizi masyarakat. Indikator ini digunakan sebagai representasi kesejahteraan pangan dan konsumsi energi penduduk.

3.2 Flowchart Proses Data Wrangling



BAB IV PEMBAHASAN

4.1 Teknik Pengambilan Data

Pada tahap ini, proses pengambilan tiga sumber data terbuka yang menyediakan informasi terkait lahan pertanian, produksi pangan, serta ketersediaan kalori per kapita. Seluruh data diunduh dalam format CSV untuk memudahkan proses pengolahan menggunakan python. Adapun rincian sumber data sebagai berikut:

1. World Bank – Agricultural Land (% of Land Area)

Data ini digunakan untuk memperoleh informasi persentase luas lahan pertanian terhadap total luas daratan suatu negara. Indikator yang digunakan adalah AG.LND.AGRI.ZS, dengan rentang tahun 1960-2021. Data ini dipilih karena memiliki tingkat cakupan negara yang luas

Sumber: *World Bank Data Catalog*

2. Our World in Data – Daily per Capita Caloric Supply

Dataset ini berisi jumlah kalori rata-rata yang tersedia per orang per hari pada suatu negara. Data ini relevan sebagai indikator ketersediaan konsumsi pangan dan sesuai untuk dianalisis bersama variabel produksi dan lahan pertanian

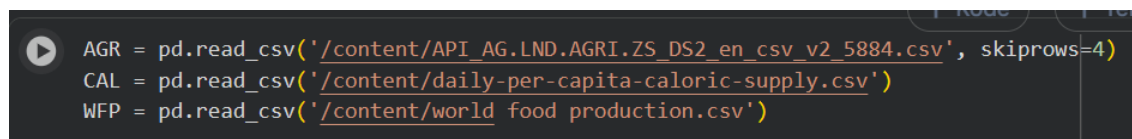
Sumber: *Our World in Data*

3. Kaggle – World Food Production Dataset

Dataset ini menyediakan data produksi berbagai komoditas pangan dengan jenis hasil pangan yang dipilih adalah beras, jagung, kopi, minyak sawit, teh, kakao dan ubi jalar. Dataset ini dipilih karena struktur datanya lengkap, terstandarisasi dan mencakup variabel produksi pangan utamanya di Asia Tenggara.

Sumber: *Kaggle Open Dataset*.

Pengambilan data dilakukan dengan mengunduh ketiga dataset menggunakan tautan yang disediakan oleh masing-masing platform. Setelah berhasil di peroleh seluruh berkas dimasukkan untuk di analisis menggunakan Python (Google Colab) dan dibaca menggunakan library pandas.



```
AGR = pd.read_csv('/content/API_AG.LND.AGRI.ZS_DS2_en_csv_v2_5884.csv', skiprows=4)
CAL = pd.read_csv('/content/daily-per-capita-caloric-supply.csv')
WFP = pd.read_csv('/content/world food production.csv')
```

Ketiga dataset tersebut kemudian diseleksi kembali untuk memastikan bahwa negara dan komponen yang dibutuhkan tersedia. Pemilahan dilakukan agar hasil analisis lebih spesifik konsisten dan dapat dibandingkan secara langsung untuk menekankan penggunaan sumber data resmi yang kredibel, rentang tahun yang konsisten, format data yang seragam, dan kesesuaian dengan tujuan proyek yaitu keterkaitan antara lahan pertanian, produksi pangan, dan ketersediaan kalori.

4.2 Data Cleaning

Data cleaning merupakan tahapan penting untuk memastikan bahwa dataset yang digunakan memiliki kualitas yang memadai sebelum dilakukan analisis lebih lanjut. Pada project ini, proses pembersihan data dilakukan terhadap tiga sumber data utama, yaitu World Food Production (Kaggle), Agricultural Land (% of land area) dari World Bank, dan Daily Per Capita Caloric Supply dari Our World in Data. Tahapan pembersihan meliputi pemeriksaan struktur data, deteksi nilai hilang, standarisasi format, serta penanganan inkonsistensi antar-dataset.

1. Pemeriksaan Struktur dan Tipe Data

Langkah pertama adalah memeriksa jumlah kolom, tipe, variabel, dan konsistensi penamaan kolom.

```
d
AGR.info()
AGR.describe()
AGR.head()
```

Tujuannya adalah untuk mengidentifikasi apakah terdapat kolom yang tidak relevan, mengetahui tipe data yang perlu dikonversi, dan menilai kelengkapan dan pola data awal.

2. Penanganan Missing Value

Setelah struktur data diperiksa, langkah berikutnya adalah menangani nilai hilang (*missing value*) yang ditemukan pada ketiga dataset. Setiap dataset memiliki karakteristik yang berbeda, sehingga disesuaikan.

➤ Dataset Agricultural Land (AGR)

```
AGR = AGR.dropna(subset=['Luas_Lahan_Pertanian_%'])
```

Nilai hilang terutama ditemukan pada tahun-tahun tertentu setelah proses transformasi bentuk data (*melt*). Karena variabel “Luas Lahan Pertanian (%)” merupakan indikator utama dan tidak dapat diestimasi dengan teknik imputasi tanpa berpotensi menghasilkan bias, maka baris yang tidak memiliki nilai pada kolom tersebut dihapus.

➤ Dataset Caloric Supply (CAL)

```
CAL = CAL.dropna()
```

Missing value relatif sedikit. Baris kosong dihapus untuk menjaga konsistensi deret waktu.

➤ Dataset World Food Production (WFP)

```
WFP[produksi_cols] = WFP[produksi_cols].fillna(0)
```

Missing value cukup banyak, terutama pada komoditas tertentu. Karena perbedaan komoditas antar-negara merupakan kondisi nyata (misalnya beberapa

negara tidak memproduksi kopi atau kakao), maka nilai hilang diganti dengan 0, bukan dihapus.

3. Standarisasi Penamaan Negara

Setiap dataset menggunakan format penamaan negara yang berbeda, misalnya “Philippines”, “Lao DPR ”, “Laos”, “Cambodia”, “East Timor”, atau “Timor-Leste”. Tanpa standarisasi, proses *merge* antar-dataset akan menimbulkan duplikasi atau ketidakcocokan data. Proses standarisasi dilakukan dengan fungsi `replace()`, seperti pada code berikut:

```
WFP["Nama Negara"] = WFP["Nama Negara"].replace({
    "Cambodia": "Kamboja",
    "Philippines": "Filipina",
    "East Timor": "Timor Leste"})
```

Standarisasi yang sama juga dilakukan pada dataset AGR dan CAL hingga seluruh dataset menggunakan satu format nama yang seragam. Daftar nama yang digunakan dalam penelitian juga disesuaikan yaitu Indonesia, kamboja, Myanmar, Malaysia, Filipina, Thailand, Timor Leste, dan Vietnam.

4. Penyelarasan Format Tahun

Dataset AGR awalnya memiliki kolom tahun dari 1960–2024. Agar bisa dianalisis sebagai deret waktu, formatnya diubah menggunakan `melt()`

```
AGR = AGR.melt(
    id_vars=['Nama Negara', 'Kode Negara'],
    value_vars=tahun_cols,
    var_name='Tahun',
    value_name='Luas_Lahan_Pertanian_%')
```

Lalu tahun dikonversi menjadi numerik dan difilter sesuai periode penelitian (2000–2021):

```
AGR = AGR[(AGR['Tahun'] >= 2000) & (AGR['Tahun'] <= 2021)]
```

5. Normalisasi dan Penyesuaian Kolom

Penyesuaian kolom dilakukan untuk memperjelas nama variabel, menyeragamkan format penulisan dan menghindari spasi berlebih atau simbol yang tidak konsisten di dataset seperti code berikut:

```
AGR = AGR.rename(columns={'Country Name': 'Nama Negara', 'Country Code': 'Kode Negara',
    'Year': 'Tahun', 'Value': 'Luas_Lahan_Pertanian_%'})
```

```
CAL = CAL.rename(columns={'Entity': 'Nama Negara', 'Code': 'Kode Negara',
    'Year': 'Tahun', 'Daily calorie supply per person': 'Kalori_per_Kapita'})
CAL["Nama Negara"] = CAL["Nama Negara"].replace({"Cambodia": "Kamboja", "Philippines": "Filipina",
    "East Timor": "Timor Leste"})
```

```
WFP["Nama Negara"] = WFP["Nama Negara"].replace({
    "Cambodia": "Kamboja",
    "Philippines": "Filipina",
    "East Timor": "Timor Leste"})
```

```
CAL.to_csv("Kalori_clean.csv", index=False)
```

Kemudian semua dataset disimpan ulang sebagai data bersih

4.3 Integrasi Data

Integrasi data dilakukan untuk menggabungkan tiga dataset yang telah melalui proses pembersihan, yaitu data luas lahan pertanian (AGR), data kalori per kapita (CAL), dan data produksi pangan (WFP). Penggabungan dilakukan berdasarkan variabel nama negara dan tahun, sehingga setiap baris data merepresentasikan kondisi satu negara pada satu tahun tertentu. Dengan sintaks sebagai berikut:

```
DATA = pd.merge(AGR, CAL, on=["Nama Negara", "Kode Negara", "Tahun"], how="outer")
DATA = pd.merge(DATA, WFP, on=["Nama Negara", "Tahun"], how="outer")
DATA = DATA.groupby("Nama Negara").apply(lambda g: g.fillna(g.median(numeric_only=True))).reset_index(drop=True)
DATA = DATA.fillna(DATA.median(numeric_only=True))
DATA = DATA.sort_values(by=["Tahun", "Nama Negara"])
DATA.to_csv("DATA_INTEGRASI_FINAL.csv", index=False)
DATA.head()
```

Proses integrasi dimulai dengan menggabungkan dataset AGR dan CAL menggunakan metode outer join, agar seluruh entri tetap dipertahankan meskipun terdapat ketidaksamaan kelengkapan data. Selanjutnya, dataset WFP ditambahkan menggunakan key yang sama. Setelah seluruh dataset tergabung, dilakukan imputasi nilai hilang menggunakan median, baik per negara maupun median keseluruhan, sehingga variabel numerik tidak lagi memiliki missing value.

Tahap akhir integrasi adalah mengurutkan data berdasarkan Tahun dan Nama Negara, kemudian menyimpan dataset hasil integrasi dalam file DATA_INTEGRASI_FINAL.csv sebagai dataset utama yang digunakan untuk analisis EDA, korelasi, regresi, dan clustering. Hasil integrasi menghasilkan satu dataset lengkap yang berisi seluruh variabel penelitian, yaitu luas lahan pertanian, delapan jenis produksi pangan, dan kalori per kapita pada periode 2000–2021 untuk delapan negara Asia Tenggara.

4.4 Analisis Data Eksplorasi

1. Statistik Deskriptif

Statistik deskriptif digunakan untuk mengetahui ringkasan data seperti nilai minimum, maksimum, kuantil, rata-rata, dan sebaran datanya. Statistik deskriptif pada python menggunakan fungsi `.describe()` dan didapatkan hasil sebagai berikut:

Index	Tahun	Luas Lahan Pertanian %	Kalori per Kapita	Produksi Jagung (ton)	Produksi Beras (ton)
count	176	176	176	176	176
Mean	2000	30,21	2579,0324	4.365.916,16	34.291.186,91
Std	6,3623	8,68	251,8545	5.839.627,09	79.000.491,64
min	2000	16,54	2043,3538	32.075	20.000
25%	2005	22,76	2372,8047	148.895,44	135.650
50%	2010	28,36	2636,3196	2.117.091	2.161.936,5
75%	2016	38,74	2798,9752	5.193.799,37	4.914.542
max	2021	45,99	3004,1028	30.253.938	318.661.730

Index	Produksi Minyak Sawit (ton)	Produksi Kopi (ton)	Produksi Kakao (ton)
count	176	176	176
Mean	204.582,56	432.240,14	499.157,77
Std	976.478,27	843.208,66	696.243,08
min	0	8	0
25%	4.143	7163,95	259,33
50%	13.402	15.722,49	116.689,2
75%	124.480	160.241,94	813.585
max	11.327.356,8	3.563.795,3	3.314.897,4

Index	Produksi Ubi Teh (ton)	Produksi Ubi Jalar (ton)
count	176	176
Mean	298.766,52	17.707.964,05
Std	354.879,73	37.758.223,62
min	168	0
25%	7392	7.026,9
50%	83.290,4	722.964,5
75%	575.411	1.376.652,25
max	1.079.779,6	131.597.530

2. Bar Chart

```
id_cols = ['Nama Negara', 'Kode Negara', 'Tahun']

num_cols = [
    'Luas_Lahan_Pertanian_%', 'Kalori_per_Kapita',
    'Produksi_Jagung_(ton)', 'Produksi_Beras_(ton)',
    'Produksi_Minyak_Sawit_(ton)', 'Produksi_Kopi_(ton)',
    'Produksi_Kakao_(ton)', 'Produksi_Teh_(ton)',
    'Produksi_Ubi_Jalar_(ton)'
]
```

Program visualisasi data dimulai dengan membuat dua list (id_cols dan num_cols) dari dataset DATA, dimana id_cols berisi kolom nama negara, kode negara, dan tahun. Sedangkan num_cols berisi list kolom luas lahan, kalori pe rkapita dan produksi bahan pangan.

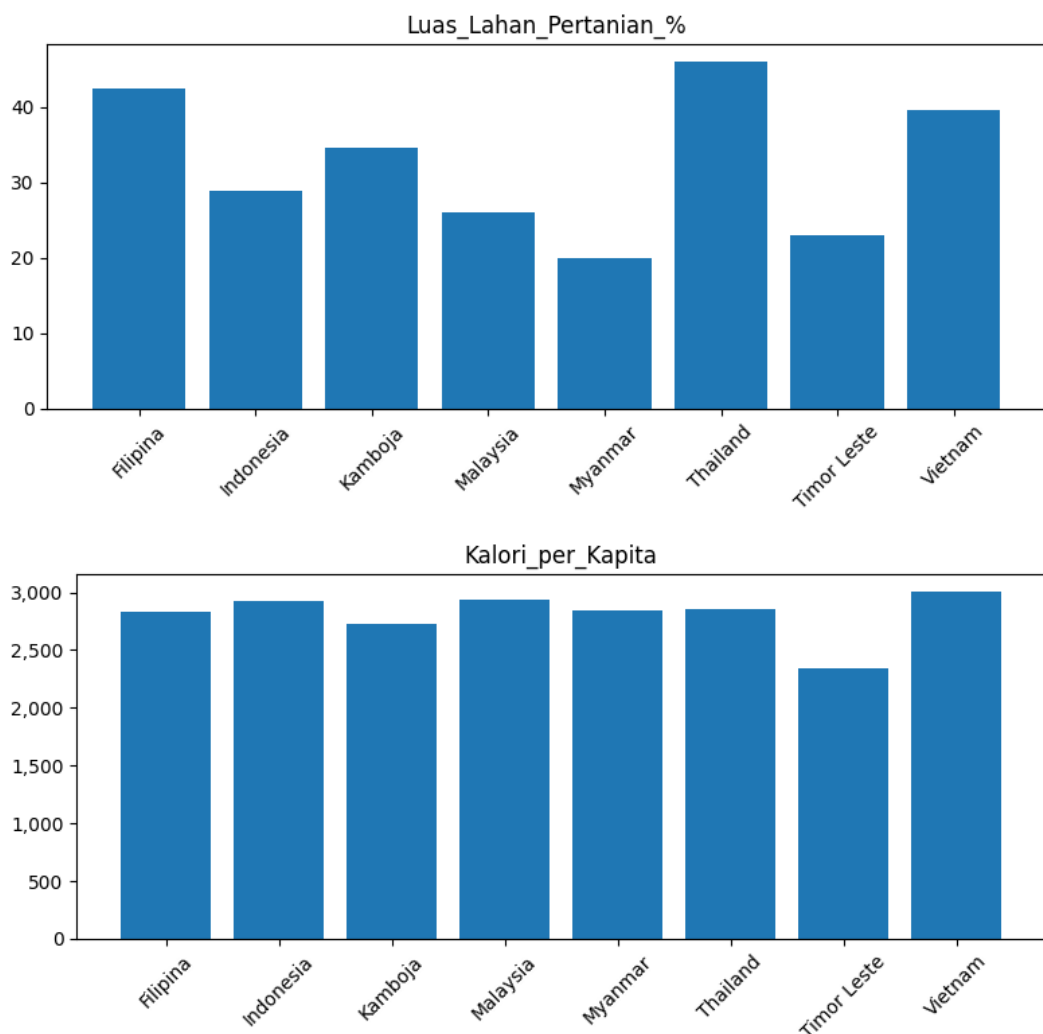
```
for col in num_cols:
    plt.figure(figsize=(8,4))
    plt.bar(DATA["Nama Negara"], DATA[col])
    plt.xticks(rotation=45)
    plt.title(col)

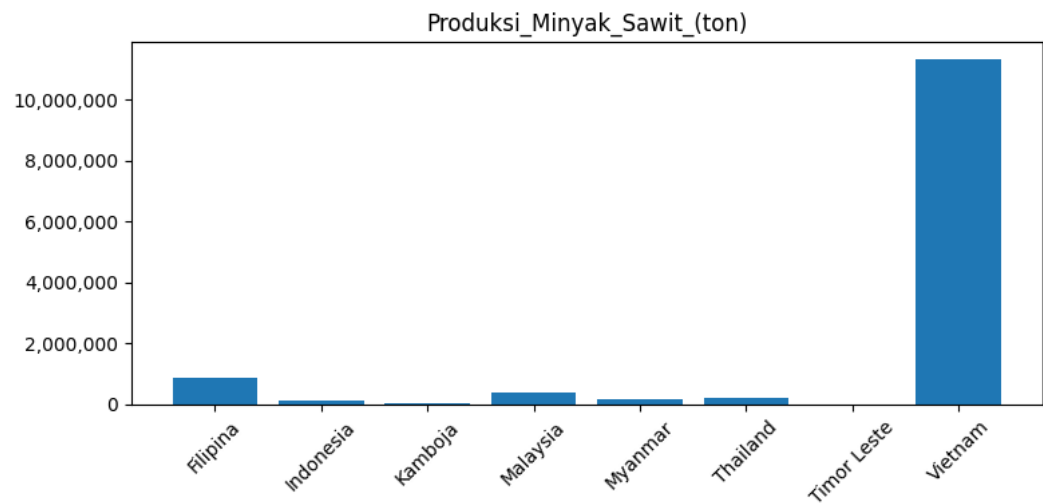
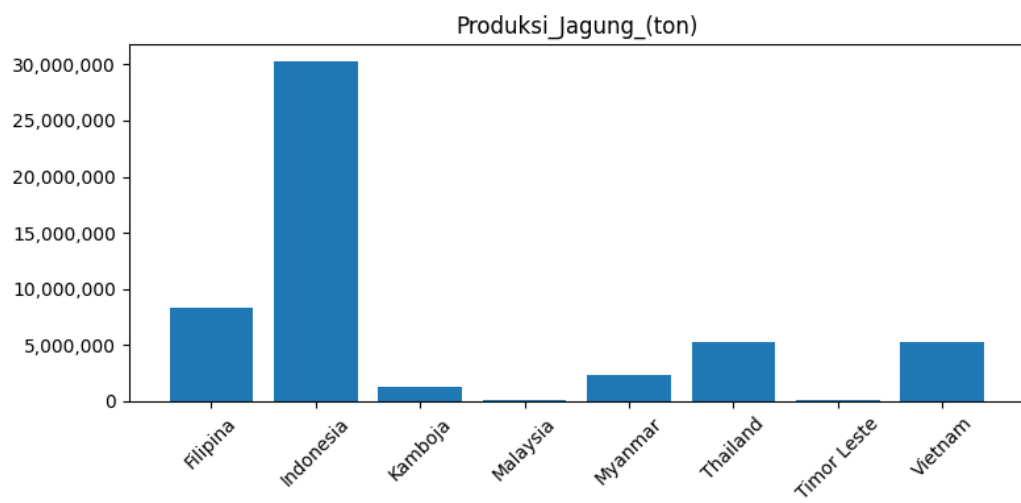
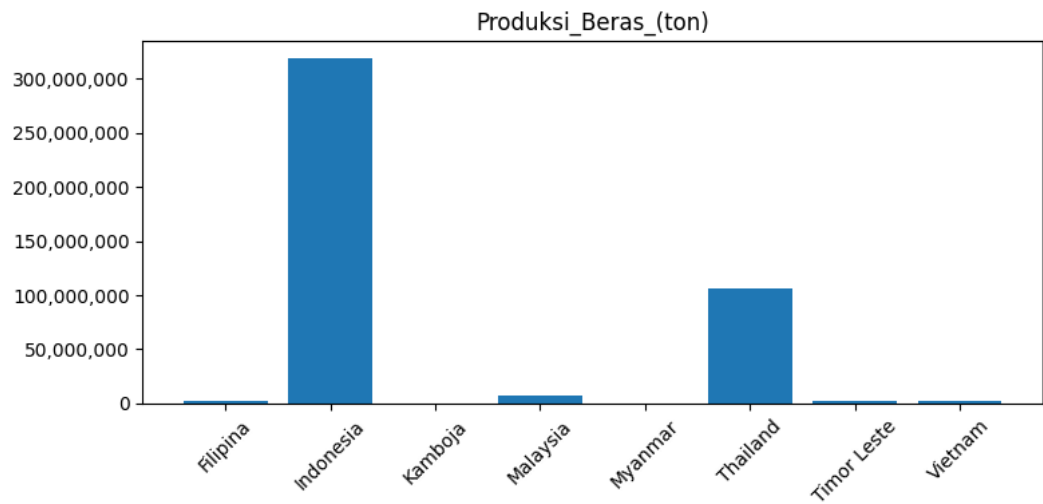
    ax = plt.gca()
    ax.yaxis.set_major_formatter(ticker.FuncFormatter(lambda x, pos: f"{int(x):,}"))

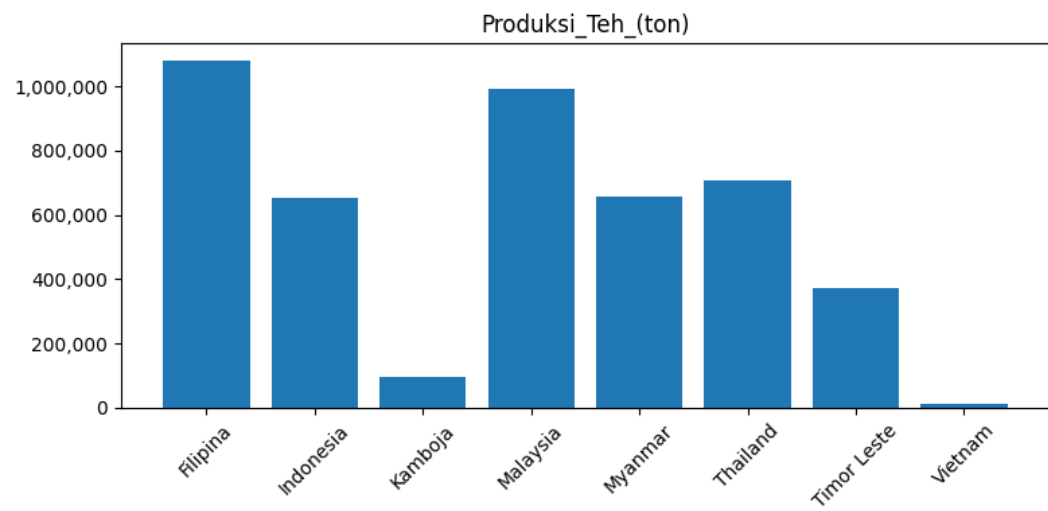
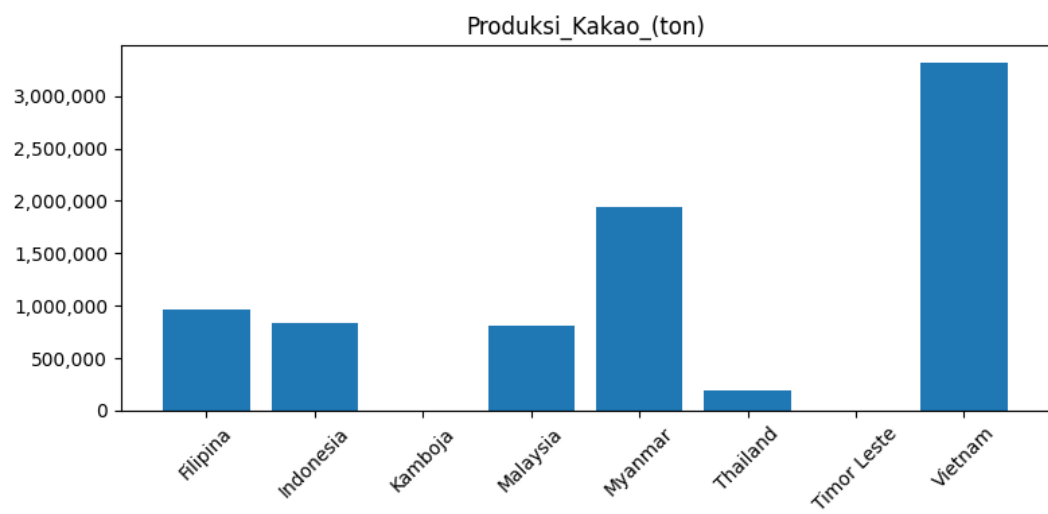
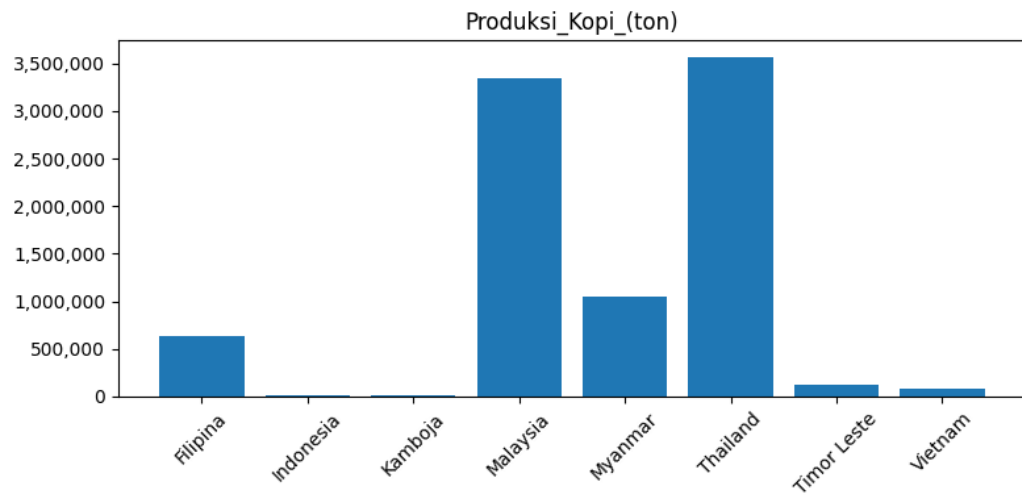
    plt.tight_layout()
    plt.show()
```

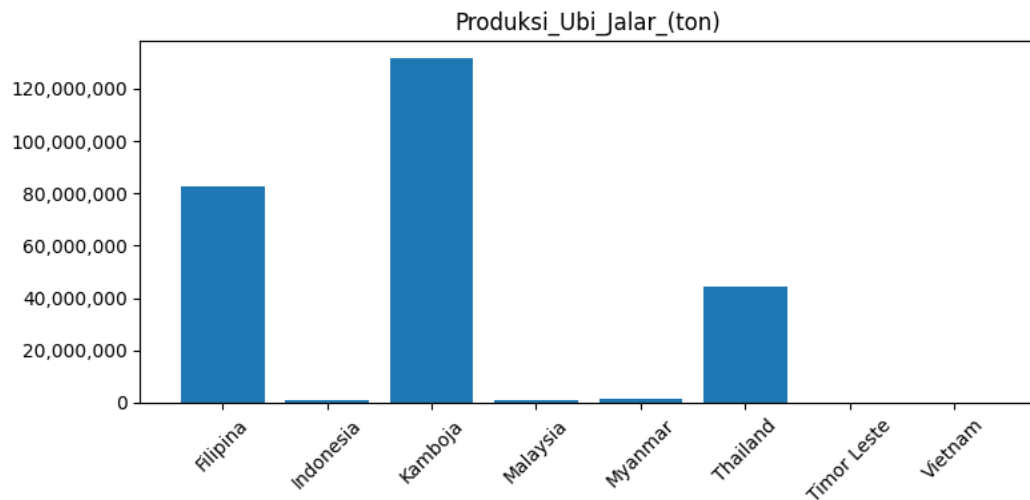
Program ini membuat diagram batang (bar chart) menggunakan library matplotlib.pyplot pada Python. Pertama, dilakukan proses looping untuk setiap kolom numerik yang terdapat pada list num_cols. Pada setiap iterasi, fungsi plt.figure

dipanggil untuk menentukan ukuran kanvas plot dengan lebar 8 inch dan tinggi 4 inch. Selanjutnya, grafik batang dibuat menggunakan fungsi `plt.bar`, di mana sumbu x berisi nama negara dan sumbu y berisi nilai dari variabel yang sedang diproses. Untuk meningkatkan keterbacaan, label pada sumbu x diputar 45 derajat menggunakan `plt.xticks`. Kemudian, judul grafik ditentukan melalui `plt.title` sesuai dengan nama kolom yang sedang divisualisasikan. Agar nilai pada sumbu y tidak tampil dalam format scientific notation, formatter khusus diterapkan menggunakan `ticker.FuncFormatter` sehingga angka ditampilkan dalam format ribuan yang lebih mudah dibaca. Setelah itu, `plt.tight_layout` dipanggil untuk memastikan tampilan plot rapi dan tidak saling bertumpuk. Terakhir, `plt.show` digunakan untuk menampilkan grafik yang telah dibuat.









3. Density Plot

```
n = len(num_cols)
rows = (n // 3) + 1
cols = 3

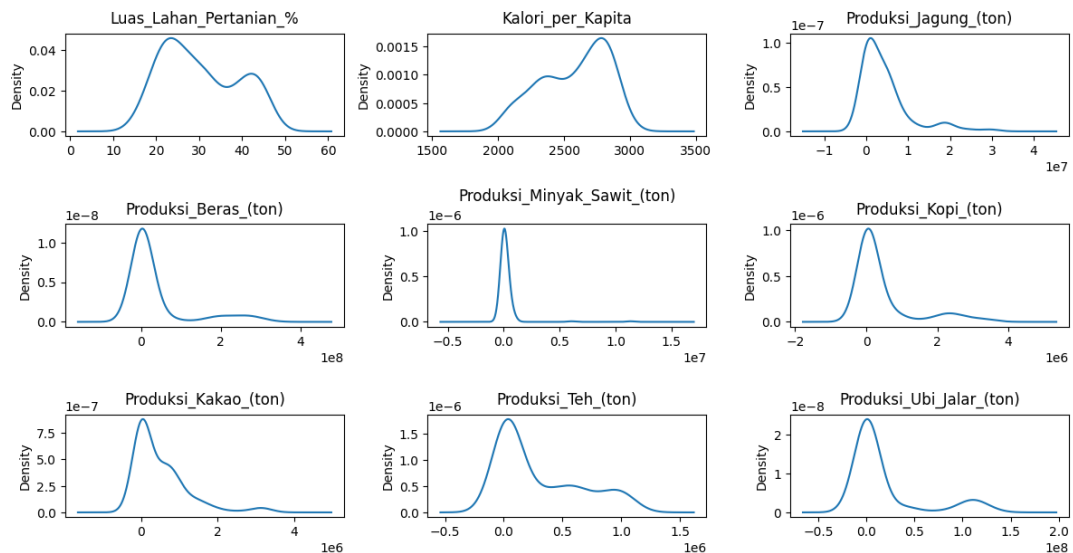
fig, axes = plt.subplots(rows, cols, figsize=(12, 8))
axes = axes.flatten()

for i, col in enumerate(num_cols):
    DATA[col].plot(kind='density', ax=axes[i])
    axes[i].set_title(col, pad=8)
    axes[i].grid(False)

for j in range(i+1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```

Program ini digunakan untuk membuat plot distribusi berupa density plot menggunakan matplotlib dan pandas. Pertama, jumlah variabel numerik pada list `num_cols` dihitung, kemudian jumlah baris pada grid subplot ditentukan dengan rumus $(n // 3) + 1$ agar seluruh plot tertata rapi dalam tiga kolom. Setelah itu, `plt.subplots` digunakan untuk membuat kanvas berukuran 12×8 inch dengan susunan subplot sesuai jumlah baris dan kolom yang telah ditentukan. Objek `axes` kemudian di-flatten agar mudah diakses menggunakan indeks. Selanjutnya dilakukan looping untuk setiap kolom dalam `num_cols`, di mana pada setiap iterasi dibuat plot densitas menggunakan fungsi `plot(kind='density')` pada axis yang sesuai. Judul subplot ditentukan menggunakan `set_title`, dan grid dinonaktifkan untuk tampilan yang lebih bersih. Setelah seluruh variabel selesai diproses, subplot kosong yang tidak terpakai dihapus menggunakan `fig.delaxes` agar hasil akhir tetap rapi. Terakhir, `plt.tight_layout` dipanggil untuk memastikan tata letak subplot tidak saling bertabrakan, dan `plt.show` digunakan untuk menampilkan seluruh grafik.



Berdasarkan output grafik yang diberikan, terlihat bahwa seluruh variabel numerik memiliki bentuk distribusi yang tidak normal. Variabel Kalori per Kapita bentuk kurva miring ke kiri (left-skewed). Sedangkan untuk variabel produksi bahan pangan seluruhnya berbentuk miring ke kanan (right-skewed). Dapat disimpulkan bahwa data belum memenuhi asumsi distribusi normal dan diperlukan transformasi.

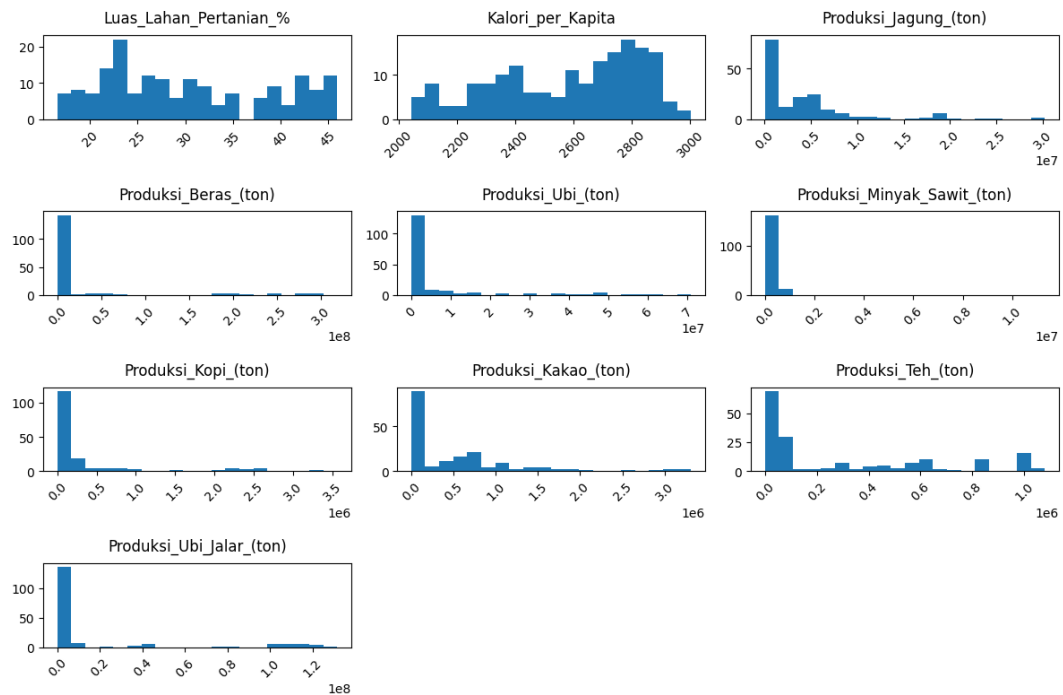
4. Histogram Distribusi Data

```
axes = DATA[num_cols].hist(bins=20, figsize=(12, 8))

for ax in axes.flatten():
    ax.set_title(ax.get_title(), pad=10)
    ax.tick_params(axis='x', rotation=45)
    ax.grid(False)

plt.tight_layout()
plt.show()
```

Program visualisasi distribusi data dimulai dengan semua kolom yang ada di list `num_cols` dibuat histogram. Hasil plot histogram tersebut disimpan dalam variabel bernama `axes`. Kemudian semua axis diratakan menggunakan `.flatten()` agar dapat diproses satu per satu di dalam loop. Pada setiap plot, judul diatur dengan jarak tambahan di atasnya (`pad=10`). Selain itu, label sumbu x diputar 45 derajat supaya tidak saling menumpuk dengan plot. Kemudian output tidak menggunakan grid supaya tampilan lebih rapi. Setelah itu, `plt.tight_layout()` dipanggil untuk memastikan tata letak antar plot tidak saling bertabrakan. Terakhir, `plt.show()` dipanggil untuk menampilkan seluruh plot histogram.



Berdasarkan output grafik yang diberikan, terlihat bahwa seluruh variabel numerik memiliki bentuk distribusi yang tidak normal. Variabel Luas Lahan Pertanian (%) berdistribusi uniform. Hal ini ditandai dengan bentuk kurva data menyebar rata. Variabel Kalori per Kapita bentuk kurva miring ke kiri (left-skewed). Sedangkan untuk variabel produksi bahan pangan seluruhnya berbentuk miring ke kanan (right-skewed). Dapat disimpulkan bahwa data belum memenuhi asumsi distribusi normal dan diperlukan transformasi.

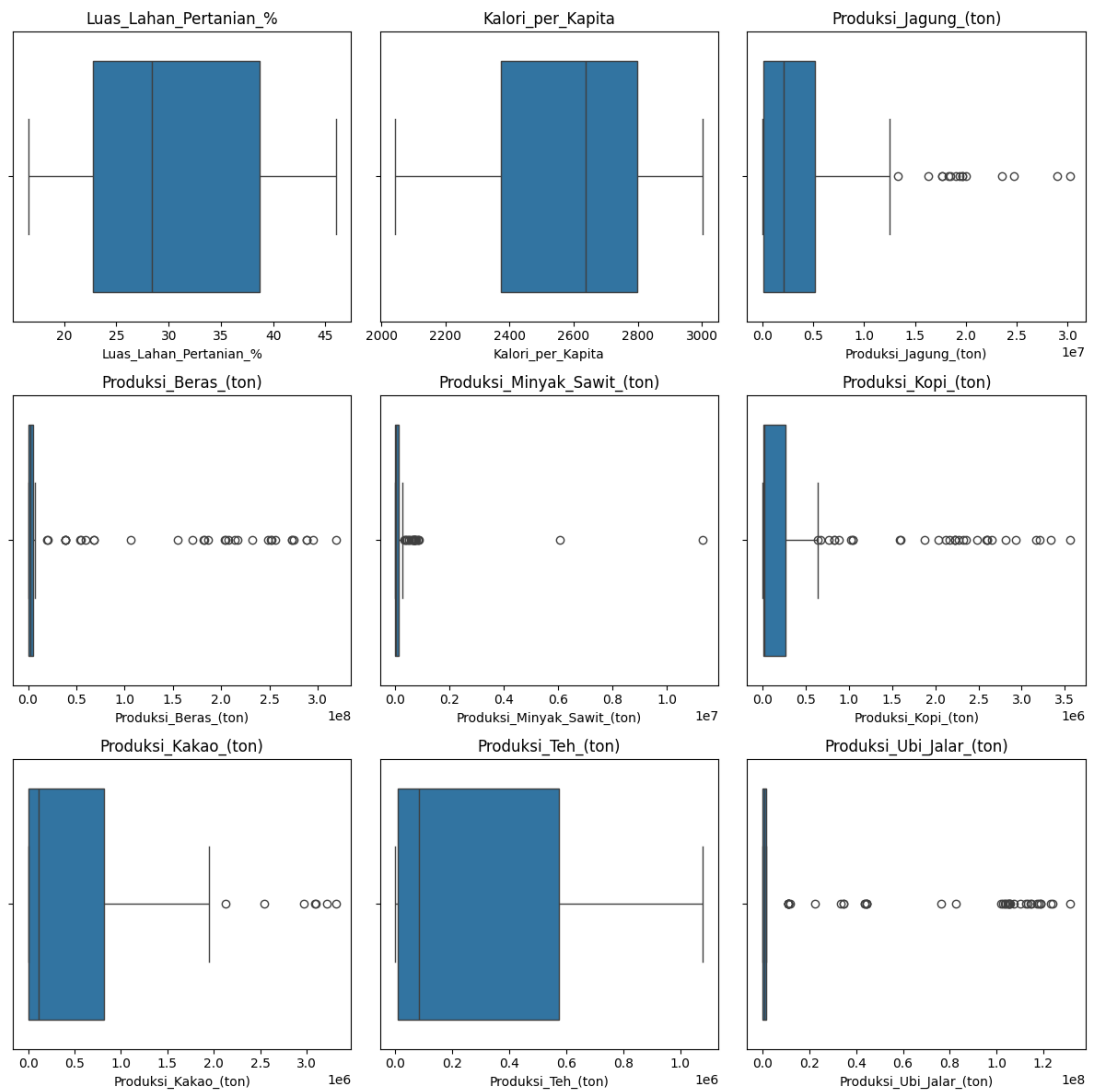
5. Box Plot

```
plt.figure(figsize=(12, 4 * ((len(num_cols) + 2) // 3)))

for i, col in enumerate(num_cols):
    plt.subplot((len(num_cols) + 2) // 3, 3, i+1)
    sns.boxplot(x = DATA[col])
    plt.title(col)

plt.tight_layout()
plt.show()
```

Program pembuatan box-plot diagram menggunakan library matplotlib.pyplot pada Python. Pertama, menggunakan fungsi plt.figure untuk menentukan ukuran kanvas plot, dimana lebarnya 12 inch dan tingginya dihitung otomatis supaya muat semua subplot, dimana tiap 3 plot tinggi bertambah 4 inch. Kemudian dilakukan looping untuk semua kolom pada list num_cols. Untuk setiap kolom, subplot disusun dalam grid dengan 3 kolom dan jumlah baris dihitung supaya semua plot rapi. Setelah posisi subplot ditentukan, boxplot dibuat menggunakan fungsi sns.boxplot. Kemudian pada plt.title() judul ditentukan sesuai nama variabel. Setelah itu, plt.tight_layout() dipanggil untuk memastikan tata letak antar plot tidak saling bertabrakan. Terakhir, plt.show() dipanggil untuk menampilkan grafik yang telah dibuat.



Berdasarkan boxplot di atas, terlihat bahwa terdapat banyak outlier pada variabel produksi jagung, produksi beras, produksi minyak sawit, produksi kopi, produksi kakao, dan produksi ubi jalar. Outlier pada data mencerminkan keragaman pada data tersebut, sehingga outlier-outlier tersebut tidak perlu dihapus karena memiliki informasi yang diperlukan.

6. Q-Q Plot

```
import scipy.stats as stats

n = len(num_cols)
rows = (n // 3) + 1
cols = 3

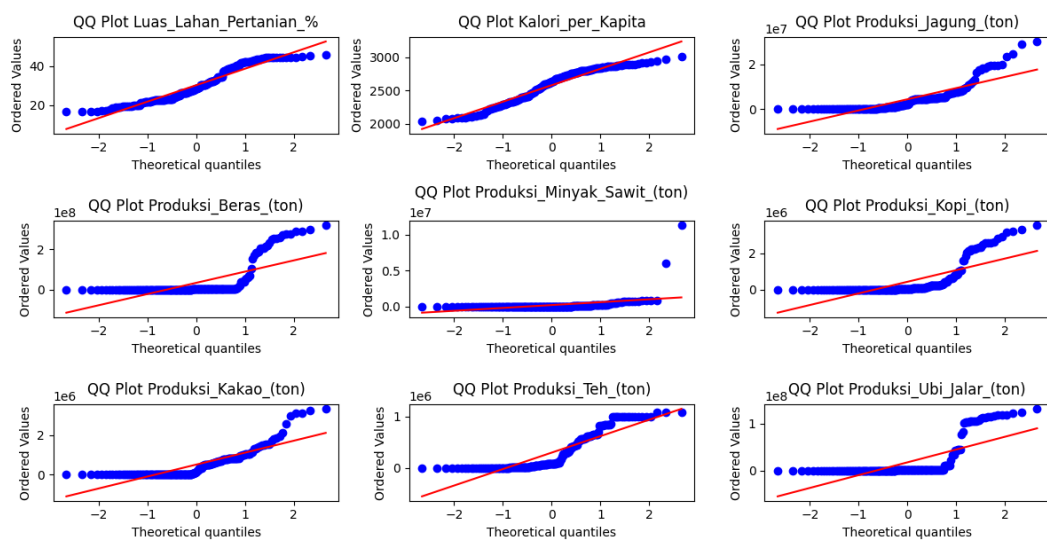
fig, axes = plt.subplots(rows, cols, figsize=(12, 8))
axes = axes.flatten()

for i, col in enumerate(num_cols):
    stats.probplot(DATA[col], dist="norm", plot=axes[i])
    axes[i].set_title(f'Q-Q Plot {col}', pad=8)
    axes[i].grid(False)

for j in range(i+1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```

Program ini digunakan untuk membuat Q-Q Plot. Pertama, jumlah variabel numerik pada list `num_cols` dihitung, kemudian jumlah baris pada grid subplot ditentukan dengan rumus $(n // 3) + 1$ agar seluruh plot tertata rapi dalam tiga kolom. Setelah itu, `plt.subplots` digunakan untuk membuat kanvas berukuran 12×8 inch dengan susunan subplot sesuai jumlah baris dan kolom yang telah ditentukan. Objek axes kemudian di-flatten agar mudah diakses menggunakan indeks. Kemudian, untuk setiap variabel, fungsi `stats.probplot` digunakan untuk membandingkan distribusi data dengan distribusi normal, sehingga terlihat apakah data mengikuti pola garis lurus atau tidak. Setiap plot diberi judul sesuai nama variabel, dan grid dimatikan biar tampilannya lebih bersih. Setelah semua plot terisi, sisa kotak kosong dihapus supaya tidak ada subplot kosong. Terakhir, `plt.tight_layout` dipakai agar jarak antar-plot tidak saling bertumpuk, lalu seluruh Q-Q Plot ditampilkan.



Q-Q Plot digunakan untuk mengecek apakah data mengikuti distribusi normal. Jika titik-titik berwarna biru mengikuti garis merah, maka dapat dikatakan bahwa data berdistribusi normal. Namun, jika banyak yang melengkung atau jauh dari garis, dapat dikatakan bahwa data tidak normal, terdapat skewness, atau ada outlier yang ekstrem. Berdasarkan Q-Q Plot, seluruh variabel menunjukkan pola penyimpangan yang cukup jelas dari garis diagonal, terutama pada upper tail. Hal ini mengindikasikan bahwa data tidak mengikuti distribusi normal dan cenderung memiliki skewness ke arah kanan dan outlier yang ekstrem.

7. Grouping dan Agregasi Data

```
output_file = "Ranking_Variabel.xlsx"

with pd.ExcelWriter(output_file, engine='openpyxl') as writer:
    for var in num_cols:
        df_rank = (
            DATA.groupby("Nama Negara")[var]
                .mean()
                .reset_index()
                .sort_values(var, ascending=False)
        )

        df_rank["Rank"] = range(1, len(df_rank) + 1)

        sheet_name = var.replace("(ton)", "").replace("%", "").replace(" ", "_")[0:30]
        df_rank.to_excel(writer, sheet_name=sheet_name, index=False)
```

Program untuk menghasilkan file Excel berisi ranking setiap variabel dimulai dengan menentukan nama file keluaran melalui variabel `output_file`. Setelah itu, `pd.ExcelWriter()` digunakan untuk membuat objek penulis Excel dengan engine `openpyxl`, sehingga setiap hasil ranking dapat disimpan ke dalam worksheet berbeda dalam satu file yang sama. Di dalam program tersebut, dilakukan iterasi terhadap setiap variabel numerik dalam `num_cols`. Untuk setiap variabel, data dikelompokkan berdasarkan Nama Negara, kemudian dihitung nilai rata-ratanya menggunakan fungsi `groupby()` dan `mean()`. Hasil tersebut di-reset index kemudian diurutkan (`sort_values`) dari nilai terbesar ke terkecil.

Selanjutnya, kolom Rank ditambahkan menggunakan fungsi `range()`, sehingga setiap negara mendapatkan peringkat berdasarkan besar kecilnya nilai rata-rata variabel tersebut. Setelah ranking terbentuk, nama sheet ditentukan melalui pembersihan teks pada nama variabel—tanda percent, satuan (ton), dan spasi dihapus, kemudian dibatasi maksimal 30 karakter agar sesuai dengan aturan Excel. Terakhir, tabel ranking yang sudah lengkap diekspor ke file Excel menggunakan fungsi `to_excel()`, dengan parameter `sheet_name` untuk menentukan nama sheet dan `index=False` untuk menghilangkan indeks bawaan Pandas. Dengan demikian, setiap variabel memiliki sheet tersendiri, dan file akhir menyatukan seluruh ranking dalam satu dokumen Excel.

Nama Negara	Kalori_per_Kapita	Rank
Malaysia	2852,365355	1
Thailand	2720,081232	2
Vietnam	2717,809432	3
Indonesia	2618,228373	4
Myanmar	2540,599936	5
Filipina	2509,47545	6
Kamboja	2476,453486	7
Timor Leste	2197,246577	8

Rata-rata kalori per kapita menunjukkan variasi antarnegara yang cukup jelas. Malaysia memiliki nilai tertinggi, yaitu sekitar 2.852 kalori, diikuti Thailand dengan rata-rata 2.720 kalori. Hal ini mengindikasikan tingkat kecukupan pangan yang relatif lebih baik di kedua negara tersebut. Sebaliknya, Timor Leste mencatatkan nilai

terendah, yaitu sekitar 2.197 kalori, yang menunjukkan konsumsi energi masyarakat yang lebih rendah dibanding negara lain di kawasan. Negara-negara lain seperti Indonesia, Vietnam, Filipina, dan Myanmar berada pada kisaran menengah.

Nama Negara	Luas_Lahan_Pertanian_%	Rank
Thailand	44,3474633	1
Filipina	40,41608356	2
Vietnam	34,535756	3
Kamboja	30,98293039	4
Indonesia	26,60674294	5
Malaysia	23,48497115	6
Timor Leste	22,72727273	7
Myanmar	18,63182437	8

Distribusi luas lahan pertanian menunjukkan bahwa Thailand dan Filipina memiliki proporsi lahan pertanian yang paling besar, masing-masing sekitar 44 persen dan 40 persen. Hal ini menandakan struktur penggunaan lahan yang masih sangat bergantung pada sektor pertanian. Sementara itu, Myanmar dan Malaysia memiliki proporsi lahan pertanian yang lebih kecil, yaitu sekitar 18 persen dan 23 persen. Indonesia, Vietnam, dan Timor Leste berada pada kisaran menengah, menunjukkan tingkat intensitas pertanian yang bervariasi.

Nama Negara	Produksi_Beras_(ton)	Rank
Indonesia	235232321,8	1
Thailand	28807518,09	2
Malaysia	4723263,091	3
Timor Leste	2459901,023	4
Vietnam	1680627,293	5
Filipina	1256572,091	6
Kamboja	96081,5	7
Myanmar	73210,45455	8

Produksi beras memperlihatkan kesenjangan yang sangat besar antarnegara. Indonesia menjadi produsen terbesar dengan rata-rata lebih dari 235 juta ton, diikuti oleh Thailand yang mencapai sekitar 28 juta ton. Negara-negara seperti Filipina, Vietnam, dan Malaysia berada pada tingkat menengah, sedangkan Kamboja, Myanmar, dan Timor Leste memiliki produksi yang jauh lebih rendah. Perbedaan ini mencerminkan kapasitas produksi dan skala pertanian padi yang berbeda-beda di setiap negara.

Nama Negara	Produksi_Jagung_(ton)	Rank
Indonesia	17349531,64	1
Filipina	6642906,795	2
Thailand	4577275,864	3
Vietnam	4236221,136	4
Myanmar	1356382,232	5
Kamboja	610203,6818	6
Timor Leste	90654,32273	7
Malaysia	64153,63909	8

Indonesia juga mendominasi produksi jagung dengan nilai rata-rata yang mencapai 17,3 juta ton, menjadikannya produsen terbesar dalam kelompok negara ini. Filipina

dan Thailand berada pada tingkat produksi menengah, sementara Malaysia, Kamboja, dan Timor Leste memiliki produksi yang jauh lebih kecil. Hal ini menunjukkan bahwa jagung merupakan komoditas strategis bagi negara-negara tertentu saja, khususnya Indonesia.

Nama Negara	Produksi_Minyak_Sawit_(ton)	Rank
Vietnam	829518,0159	1
Filipina	569332,6182	2
Thailand	141388,7173	3
Malaysia	35684,95455	4
Myanmar	26922,04227	5
Indonesia	23032,62727	6
Kamboja	7480,545455	7
Timor Leste	3301	8

Dalam produksi minyak sawit, Malaysia menempati posisi teratas dengan rata-rata mencapai 1,9 juta ton. Kontribusi ini jauh melampaui negara lain dalam sampel, sehingga menunjukkan dominasi Malaysia pada sektor sawit. Thailand, Myanmar, dan Filipina memiliki produksi yang lebih rendah namun tetap signifikan. Sementara itu, Indonesia pada dataset ini terlihat memiliki nilai yang relatif kecil, kemungkinan dipengaruhi oleh ketersediaan data yang terbatas dalam periode tertentu.

Nama Negara	Produksi_Kopi_(ton)	Rank
Malaysia	1918060,636	1
Thailand	973139,1481	2
Myanmar	356993,8348	3
Filipina	90899,28291	4
Timor Leste	89062,90909	5
Vietnam	10637,90127	6
Indonesia	9911,623182	7
Kamboja	9215,784545	8

Malaysia dan Thailand merupakan dua negara dengan produksi kopi tertinggi, masing-masing sekitar 1,9 juta ton dan 973 ribu ton. Filipina, Myanmar, dan Indonesia berada pada level menengah. Sebaliknya, Vietnam dan Timor Leste menunjukkan produksi kopi yang sangat rendah dalam dataset ini. Perbedaan ini dapat mencerminkan kondisi agroklimat serta fokus komoditas unggulan tiap negara.

Nama Negara	Produksi_Kakao_(ton)	Rank
Vietnam	1744029,464	1
Myanmar	762898,0114	2
Filipina	654788,0953	3
Indonesia	603365,2218	4
Malaysia	198262,6528	5
Thailand	29874,41164	6
Timor Leste	29,37409091	7
Kamboja	14,96636364	8

Untuk komoditas kakao, Vietnam mencatatkan produksi tertinggi dengan rata-rata lebih dari 1,7 juta ton, diikuti Myanmar dengan sekitar 762 ribu ton. Malaysia dan Filipina memiliki nilai menengah, sedangkan Kamboja dan Timor Leste menunjukkan produksi yang sangat kecil. Variasi ini mengindikasikan bahwa budidaya kakao hanya menjadi komoditas utama di beberapa negara tertentu.

Nama Negara	Produksi_Teh_(ton)	Rank
Malaysia	815913,1727	1
Filipina	614954,0091	2
Myanmar	385870,4636	3
Indonesia	351700,6182	4
Thailand	155680,0455	5
Timor Leste	49582,15455	6
Kamboja	8822,409091	7
Vietnam	7609,363636	8

Produksi teh tertinggi terdapat pada Malaysia dan Filipina, masing-masing sekitar 815 ribu ton dan 614 ribu ton. Myanmar dan Indonesia mencatatkan nilai menengah, sementara Timor Leste dan Vietnam memiliki nilai terendah. Distribusi ini menggambarkan bahwa teh bukan merupakan komoditas utama di sebagian besar negara dalam dataset ini.

Nama Negara	Produksi_Ubi_Jalar_(ton)	Rank
Kamboja	111999643,6	1
Thailand	19193984,2	2
Filipina	8225621,273	3
Myanmar	1085607,136	4
Indonesia	625950,8209	5
Malaysia	529793,0768	6
Timor Leste	1665,454545	7
Vietnam	1446,885273	8

Produksi ubi jalar menunjukkan perbedaan yang sangat mencolok antarnegara. Kamboja menjadi negara dengan produksi tertinggi, mencapai rata-rata lebih dari 111 juta ton, jauh melebihi negara lain. Thailand berada di posisi kedua dengan sekitar 19 juta ton. Filipina, Indonesia, Malaysia, dan Myanmar mencatatkan produksi dalam jumlah menengah, sementara Vietnam dan Timor Leste memiliki nilai yang paling rendah. Hal ini menunjukkan bahwa ubi jalar merupakan komoditas utama yang sangat menonjol di Kamboja.

8. Pivoting

```
output_path = "pivot_per_variabel.xlsx"
with pd.ExcelWriter(output_path, engine='openpyxl') as writer:

    for col in num_cols:
        pivot_df = DATA.pivot_table(
            index='Nama Negara',
            columns='Tahun',
            values=col
        )

        sheet_name = col.replace("(ton)", "").replace("%", "").replace(" ", "_")[0:31]
        pivot_df.to_excel(writer, sheet_name=sheet_name)
```

Program untuk menghasilkan output pivot setiap variabel dimulai dengan menentukan lokasi penyimpanan file menggunakan variabel `output_path`. Selanjutnya, `pd.ExcelWriter` dengan engine `openpyxl` digunakan untuk membuat file Excel baru yang dapat berisi banyak sheet dalam satu file. Di dalam `with`, proses dilakukan untuk setiap variabel numerik yang tercantum dalam `num_cols`. Untuk setiap variabel, fungsi `pivot_table` digunakan untuk membuat tabel ringkas dengan baris berupa Nama Negara, kolom berupa Tahun, dan nilai berupa data numerik dari variabel tersebut. Hasil pivot ini kemudian disimpan pada variabel `pivot_df`. Setelah tabel pivot terbentuk, nama sheet disesuaikan dengan melakukan pembersihan teks, seperti menghapus simbol `"(ton)"`, persen, dan spasi, kemudian memotong panjang nama sheet agar tidak melebihi batas 31 karakter. Terakhir, setiap tabel pivot dituliskan ke dalam file Excel menggunakan `to_excel`, sehingga setiap variabel memiliki sheet tersendiri yang berisi ringkasan lengkap per negara per tahun. Didapatkan hasil excel sebagai berikut:

- Data runtun waktu Luas Lahan Pertanian

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Filipina	37,67649	37,5088	37,34447	37,69662	38,46799	38,0823	38,83691	39,5412	40,27904	40,58088	40,58088	41,11748
Indonesia	23,76434	24,20268	24,52226	24,78856	24,83737	24,93943	25,52086	26,04902	26,20434	26,57272	26,56826	27,09642
Kamboja	27,02243	27,70224	28,3254	28,89191	29,62837	30,34217	30,33651	30,33651	30,1877	30,48613	30,77949	
Malaysia	21,11916	21,14442	21,41988	21,65953	21,70659	21,71484	21,53462	21,63202	21,76351	21,97535	22,49064	22,92071
Myanmar	16,54375	16,73807	16,71665	16,80387	16,96	17,23858	17,7468	18,33762	18,792	19,04332	19,1746	19,2227
Thailand	42,85658	43,03862	43,22065	43,40269	43,58472	43,76676	43,9488	44,12887	44,31091	44,49294	44,67498	44,68281
Timor Leste	22,66308	22,63618	22,60928	22,44788	22,42098	22,39408	22,36718	22,34028	22,64963	22,75723	22,73033	22,77068
Vietnam	28,22607	30,48314	30,44598	30,75757	31,59158	32,13846	32,21919	32,19434	32,73997	32,90681	34,34506	34,37328

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
	41,68763	41,81843	41,91569	42,01295	42,11021	42,22423	42,30137	42,37851	42,45565	42,53614
	27,09196	27,35381	27,50914	27,61121	27,87306	27,75989	28,41183	28,75033	28,96499	28,95585
	31,1509	31,45244	31,68261	31,99606	32,29098	32,77036	33,18811	33,71628	34,30139	34,69998
	23,82955	24,5299	24,6203	26,08422	26,08735	26,08735	26,08735	26,08735	26,08735	26,08735
	19,21304	19,27329	19,34219	19,52134	19,54717	19,99786	20,02651	19,88571	19,88754	19,88754
	44,69064	44,69899	44,71021	44,71804	44,71021	44,72	44,71745	45,01948	45,55168	45,99816
	22,8043	22,83793	22,87155	22,90518	22,9388	22,97243	23,00605	22,95898	22,95898	22,95898
	34,43311	34,67267	34,69913	38,76343	38,84079	38,82796	38,73445	39,5241	39,43477	39,43477

- Data runtun waktu Kalori per Kapita

ama Negai	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	2332,718	2311,679	2332,075	2330,984	2399,371	2410,353	2453,839	2507,943	2518,129	2526,313	2411,137
Indonesia	2355,519	2345,14	2373,19	2392,309	2399,094	2399,685	2416,755	2457,218	2481,343	2541,644	2574,33
Kamboja	2074,381	2140,174	2238,243	2240,734	2258,556	2312,594	2362,355	2371,65	2383,767	2394,514	2521,02
Malaysia	2897,056	2866,865	2809,854	2817,932	2840,118	2792,748	2799,334	2818,618	2759,903	2805,206	2833,503
Myanmar	2043,354	2090,522	2133,136	2185,259	2217,784	2260,55	2332,388	2410,216	2508,426	2547,319	2688,466
Thailand	2576,524	2554,422	2610,83	2626,516	2636,214	2701,113	2767,284	2723,733	2698,499	2659,989	2742,888
Timor Leste	2145,616	2080,313	2109,532	2094,192	2123,26	2133,138	2124,373	2109,325	2096,072	2050,052	2259,181
Vietnam	2347,36	2409,857	2473,121	2531,062	2595,579	2585,473	2591,186	2636,95	2682,334	2691,089	2729,502

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
2418,934	2449,567	2444,127	2463,227	2510,828	2592,294	2636,425	2728,61	2801,073	2793,962	2834,871
2629,694	2672,883	2711,15	2786,576	2764,305	2857,988	2930,755	2886,825	2875,656	2880,005	2868,959
2610,723	2602,176	2590,716	2603,789	2642,222	2651,58	2675,357	2674,9	2716,717	2721,388	2694,421
2865,85	2888,144	2847,142	2833,239	2876,509	2864,132	2893,548	2893,876	2893,285	2937,165	2918,009
2678,092	2688,896	2704,655	2748,753	2775,75	2794,168	2808,309	2798,856	2795,324	2845,753	2837,224
2730,821	2752,722	2722,841	2760,611	2775,562	2763,117	2751,862	2777,046	2800,525	2850,377	2858,291
2289,795	2280,19	2288,591	2339,49	2299,941	2286,182	2309,426	2249,453	2257,044	2222,231	2192,031
2757,689	2815,859	2800,417	2833,538	2822,843	2847,281	2862,369	2885,865	2922,106	2966,226	3004,103

- Data runtun waktu Produksi Jagung (ton)

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	4511104	4525012	4319262	4615625	5413386	5253160	6082109	6736940	6928225	7034033	6376796
Indonesia	9677000	9347200	9585277	10886442	11225243	12523894	11609463	13287527	16323922	17629740	18327636
Kamboja	156972	185589	148897	314591	256665	247760	376938	523000	611865	924026	773269
Malaysia	65000	67000	70000	72000	72000	75000	80000	32075	32959	36396	47602
Myanmar	358900	524000	593400	692900	771100	903500	1015800	1128100	1184700	1225700	1354400
Thailand	4472903	4496960	4259289	4248989	4341474	4093634	3918332	3890218	4249354	4616119	4860746
Timor Leste	88449	69000	93714	70175	82209	92219	118984	71526	100173	134715	148890,8
Vietnam	2005900	2161700	2511200	3136300	3430900	3787100	3854500	4303200	4573100	4371700	4606800

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
6971221	7407069	7377294	7770603	7518756	7218817	7914909	7771919	7978845	8118546	8300322
17643250	19387022	18511852	19008426	19612436	23578412	28924016	30253938	19650000	24687000	20010000
717000	950909	927000	550000	400000	663000	750000	1232000	895000	900000	920000
59842	83601	86499	59188	62460	64867	72560,52	68906,65	60497,87	68207,37	74718,65
1461500	1501600	1600600	1693400	1748500	1830631	1909295	1984137	1985765	2072482	2300000
4972952	4947530	4876180	4804670	4729527	4390185	4820962	5069143	4535058	4805844	5300000
49783	62839	100789	102473	142361	96565,16	75000	85176,16	75690	77606	56058
4835717	4973482	5190896	5202511	5287261	5244141	5109766	4874054	4732149	4558107	4446383

- Data runtun waktu Produksi Beras

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	1467138	1558048	1363460	1457175	1578169	1451011	1666273	1700229	1589504	1732210	1581732
Indonesia	170547800	154729570	181787630	186815100	182122600	203381380	204825490	217056820	214502800	207117860	2,32E+08
Kamboja	80440	79956	76700	84000	93488	36719	76234	67100	86284	54953	34350
Malaysia	6933328,5	7328227,5	5593342	4403684,5	5718187	6004940	5787142,5	5517663,5	4913542	5390533,5	5050676
Myanmar	60281	68372	65177	66500	45000	50000	53000	68500	37400	46600	57900
Thailand	67833279	67833279	106655105	1754911,2	19125019	474156,29	458664,09	1823877,4	1754911,2	4518190,4	4498566
Timor Leste	2222693	2093997	2167715	2085253	2369446	2264663	2236081,5	2037851,9	2253098,5	2156158	2711356
Vietnam	209600	243317,41	264422,21	354891,01	651298	1015513,8	1399368,6	1792882,4	2219230,2	2243870,2	2297200

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
1587776	1754049	1949796	1791211	1750729	1560870	20184	22700	20000	21600	20722
2,49E+08	2,52E+08	2,52E+08	2,56E+08	2,73E+08	2,73E+08	2,75E+08	2,88E+08	2,88E+08	2,95E+08	3,19E+08
76380	97640	120010	104790	181170	95440	99740	131180	115560	165010	156649
3217061	4901207	4904148	4917542	2458000	2996000	3187000	3319000	3460000	3856422	4054142
60100	38700	47700	45700	82800	110000	137000	152600	115500	131600	70200
54908965	3179713	4498566	4498566	20446043	38658103	38973514	38973514	38763802	59919636	54215018
2755966	2678006	2722871	2487020	2575901	2635004	2709424	2888461	2674145	2612330	2780382
2348009	1941002	2500849	2455972	2550979	2471052	2505111	1993271	1550417	1975222	1990322

- Data runtun waktu Minyak Sawit (ton)

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	248424,6	866277,4	548725	690041,8	370442	412074,2	846490,4	721902,2	721902,2	656869,2	780156,4
Indonesia	43284,4	40400	7520	7520	7520	7980	126520	13000	13000	40600	9297,4
Kamboja	0	0	0	0	0	0	0	0	2138	2333	3581
Malaysia	17045,2	17045,2	17045,2	13227,8	17045,2	17045,2	17045,2	17045,2	17045,2	17045,2	17045,2
Myanmar	24800	23991,53	10212,8	10152,8	40866,4	14391,6	17406,4	17406,4	14391,6	5581	8022
Thailand	93600	93600	55350	13402	134602	13402	20268	20268	13402	214714	214714
Timor Leste	3700	3955	5376	5728	3800	4524	2400	2570	2234	2216	2750
Vietnam	149523,6	3140	3080	3080	2434,8	2434,8	3191,726	2988,272	2320	27138,2	3472,4

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
690041,8	656869,2	682565,6	838922,4	490986,8	638357,4	638357,4	511287	248424,6	142400	123800
9297,4	9297,4	9297,4	9297,4	9297,4	9297,4	9297,4	19731,4	9297,4	32031,4	63934
6600	7920	5800	6200	8000	10000	13000	16000	22000	25400	35600
4623,4	17045,2	17045,2	17045,2	404354	100583,4	4623,4	6840,8	6840,8	9262,6	13125,2
192800	8004,6	36108,8	36076,4	8055,6	13156,4	13156,4	38585,6	8022	11205,4	39891,2
214714	214714	214714	214714	214714	134602	214714	214714	156202	214714	214714
2913	3650	2760	2800	2770	2485	2584	3135	3128	3344	3800
3472,4	4012,4	4143	4143	4143	4143	146959	146959	342953,8	6058307	11327357

- Data runtun waktu Produksi Kopi (ton)

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	640614,2	881	1719,4	1105,4	501727	565572,6	853,2	1148,6	1148,6	1026,2	1102,8
Indonesia	12927	12701	12701	15478	12700	12701	12175	11567	13393	7000	8250
Kamboja	8000	8000	8000	8000	8000	8000	8000	8000	8000	8000	8631
Malaysia	2200	2490	2320	1960	2270	3163825	3213383	2602952	2032553	3341727	2269018
Myanmar	669947,4	281519,6	637494,7	639672,8	414619,8	268,2	205647,9	205604,7	307,654	200889,6	411943,9
Thailand	420929,4	420929,4	770794,6	828703,3	2226347	2933242	3563795	2224491	828703,3	57043,98	241084,4
Timor Leste	58300	59500	62048	71190	70103	66152	80303	101218	84332	75082	91334
Vietnam	84341,6	33980,8	35540,8	35513,46	26750,99	15966,99	184,242	170,626	159,766	158,92	135,64
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	
1105,4	1026,2	1112,2	986,6	4548,6	897,6	897,6	4477,6	252703,8	7218,6	7911	
8650	7600	9308	10206	7938	8000	8100	7263,3	6928,77	6431,97	6036,67	
8538,26	7734	8000	9000	10000	10987	9723	11000	12000	13000	14134	
2590240	2162567	2351421	1874104	2658582	2598195	2122420	2809096	2484059	1585004	2326948	
8	211334,4	211334,4	211267,6	210776,9	342339,3	1047265	7805,8	12527	1047265	884025,3	
241084,4	241084,4	241084,4	241084,4	1021361	2226347	287959,6	287959,6	1602056	253149,4	249828,6	
99717	87436	86064	118500	93600	113926	104679	128700	116900	103900	86400	
126	133	141	137	147	150	84	106	58	24	24	

- Data runtun waktu Produksi Kakao (ton)

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	315480	965213,6	638500,8	765540,6	485136,1	521469,2	940421,8	777166,4	777166,4	815467,6	824028,6
Indonesia	829075,9	828252,9	821955,9	774708,1	719504,8	658298,8	658518,8	597062,9	537959,1	474917,1	489095,8
Kamboja	15	5	16	15	12	15	12	14	13	5	4
Malaysia	6520	6520	6520	5066	6520	6520	6520	6520	6060	6520	6060
Myanmar	1613,2	6082,886	7707,632	7731,632	559697,4	535031,6	1016309	1522641	28140	1009713	1013445
Thailand	110589,8	110589,8	189228,4	404,712	1180,4	67905,71	67905,71	94324,6	404,712	351,012	319
Timor Leste	33,06	33,03	33	32,97	32,62	35	33,04	32,91	32,78	32,64	32,51
Vietnam	122788,6	239231,4	245769	714534,4	1797270	1824167	3096883	3092753	3215576	3314897	2971248
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	
765540,6	815467,6	783157,4	966033,2	640758,8	813585	813585	655739	315691,2	5134,6	5054,6	
448596,4	474443,3	493260,1	477518,6	499848,9	495428,1	516334	571531,3	596829	686146,8	624748,4	
1	26	23	8,92	7	3,4	12	16	16,5	9,1	80,34	
332743,7	6060	6060	6060	815213,2	815213,2	815213,2	815213,2	335990,6	339598,6	5066,608	
31226,2	1020727	1020727	1019833	1056760	670630,8	1142688	524941,8	1499262	1142688	1946161	
319	319	319	899,8	1180,4	1577,6	1577,6	2086,4	1834	1834	2086,4	
30	31,24	32,12	31,92	31,65	31,62	31,48	31,37	31,27	0	0	
2545430	2130210	1290820	1717648	1322427	1351043	1447641	1388746	1471684	1533941	1533941	

- Data runtun waktu Teh

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	3592,4	837345,2	1079779,6	1079779,6	245883,8	818118,8	854811,6	837345,2	837345,2	818118,8	837345,2
Indonesia	612824,4	130440,8	637720,2	652913,4	644292,8	587532,2	587532,2	618048,6	618048,6	618048,6	618048,6
Kamboja	270	245	168	232	200	200	200	200	200	200	200
Malaysia	992570,2	992570,2	992570,2	993490,4	992570,2	992570,2	992570,2	992570,2	993490,4	992570,2	993490,4
Myanmar	635747,4	321972,2	571370,6	570247,2	658347	81015,6	267408,8	267408,8	81015,6	214375	416203
Thailand	645123,4	645123,4	707312,8	51469,6	69967	51469,6	26395,8	26395,8	51469,6	83290,4	83290,4
Timor Leste	5182,6	5789,6	5789,6	5789,6	5789,6	5789,6	39669,4	31697	31697	22480,8	51147
Vietnam	4383	5007,4	2339	3158,4	3928,6	3928,6	4363,2	6317	6325,6	7327,2	8528,6
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	
1079779,6	818118,8	818118,8	837345,2	3592,4	526698,8	526698,8	473753,2	3592,4	97128,2	94696,6	
302506,8	302506,8	302506,8	150023,8	330	16485,2	31449,8	67082,8	71092,4	71092,4	96886,4	
200	300	350	400	500	500	500	400	254	92435	95939	
562956,8	993490,4	993490,4	993490,4	377,2	331,2	371785	563198	562956,8	993490,4	993490,4	
297405,6	415963	292894,4	292894,4	427221,6	440784,2	440784,2	505509,4	416203	433595	440784,2	
83290,4	83290,4	83290,4	83290,4	83290,4	69967	83290,4	83290,4	164072,6	83290,4	83290,4	
31697	41400,8	55651,4	55651,4	55651,4	67749,2	69020,4	67749,2	58183	370856	6375,8	
8528,6	9319,4	11617,6	11617,6	11617,6	11617,6	11617,6	11617,6	9460,2	7392,8	7392,8	

- Data runtun waktu Produksi Ubi Jalar (ton)

Nama Negara	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Filipina	580553	720804	730830	713500	743937	819641	836560	883100	1081592	1223068	1201200
Indonesia	548650,3	521326,9	533677,1	529145,4	538315	541057	559713,8	562862,06	575684,94	579619,9	593606,6
Kamboja	112551736	123346060	131597530	117911170	118845550	106014160	110291840	119020370	114605784	1,04E+08	1,02E+08
Malaysia	292116	278293	288164	269670,44	256434	275322,8	300987,78	243157,45	305729,22	738520	695467
Myanmar	834853	914488	965613	993794	1002708	1027946	1010738	1037873	1057125	1069687	1115724
Thailand	34524863	34524863	11786182	44315492	33496209	44286045	43744089	43780835	44315492	22330862	10813932
Timor Leste	3000	3000	3000	2000	1040	1000	1000	1000	1100	1300	1300
Vietnam	0	0	0	7209,2	100	200	300	400	500	500	500

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
1124230	1341910	1406860	1430903	1616880	1241850	1450249	1366583	1308728	76507740	82632950
613883,5	633785,7	651661	673115	689606	696583	707551,2	729834,6	745543,5	765605,6	780090
1,15E+08	1,03E+08	1,06E+08	1,05E+08	1,06E+08	1,05E+08	1,13E+08	1,08E+08	1,17E+08	1,24E+08	1,03E+08
748802	783647	725125	701237	686761	672395	706394	731394	656717	651468	647646
1119814	1128836	1182524	1190464	1215062	1232280	1265245	1187860	1133860	1105680	1091183
10788497	10823038	10813932	10813932	10796853	6480	74624,12	74624,12	39134,12	58865,92	58805,92
1200	1300	1300	1300	1500	1500	1800	1800	1900	2000	2100
500	500	800	700	900	1000	1000	3770,124	6361,138	4391,014	2200

Pivoting tersebut dilakukan untuk mengubah struktur data menjadi format yang lebih terorganisir sehingga setiap negara memiliki deret waktu yang runtut dan mudah dibandingkan. Dengan format tabel yang sudah rapi ini, pola perubahan tiap variabel dari tahun ke tahun dapat diamati dengan lebih jelas. Setelah data berada dalam bentuk time-series yang siap dianalisis, langkah selanjutnya adalah menampilkan tren tersebut menggunakan visualisasi line plot untuk melihat dinamika dan perkembangan setiap negara sepanjang waktu.

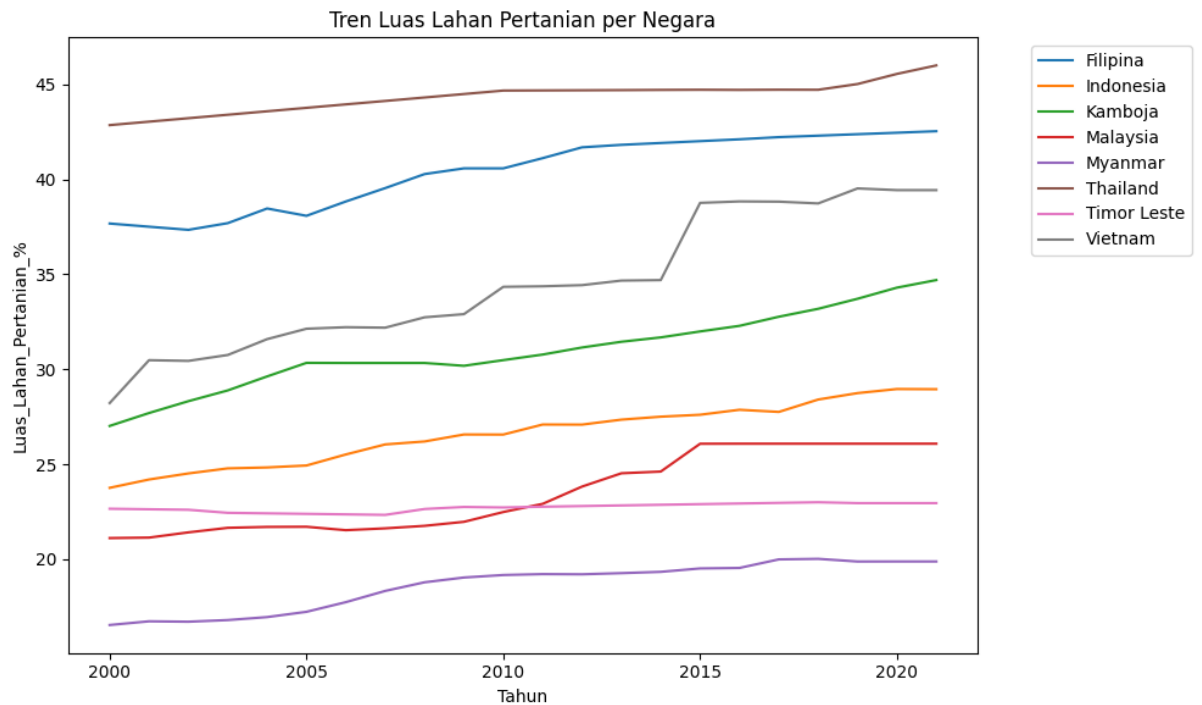
9. Visualisasi Tren Luas Lahan Pertanian (%)

```
plt.figure(figsize=(10,6))

sns.lineplot(
    data=DATA,
    x='Tahun',
    y='Luas_Lahan_Pertanian_%',
    hue='Nama Negara'
)

plt.title('Tren Luas Lahan Pertanian per Negara')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.savefig('tren_lahan.pdf', format='pdf')
```

Program visualisasi tren luas lahan pertanian dimulai dengan membuat figure berukuran 10x6. Selanjutnya, fungsi lineplot dari seaborn digunakan, dimana dataset yang digunakan adalah DATA, sumbu x adalah tahun, sumbu y adalah persentase luas lahan pertanian, dan setiap negara diberi warna yang berbeda. Kemudian dengan plt.title() untuk membuat judul grafik. Selain itu, diatur juga posisi legenda grafik supaya tidak berada di area kurva, fungsi bbox_to_anchor digunakan untuk menentukan titik koordinat legenda, dan fungsi loc untuk menentukan posisi relatifnya. Setelah itu, plt.tight_layout() dipanggil untuk memastikan tata letak antar plot tidak saling bertabrakan. Terakhir, plt.savefig digunakan untuk menyimpan hasil plot dalam format pdf



Grafik di atas menunjukkan bahwa negara-negara di Asia Tenggara mengalami tren kenaikan luas lahan pertanian dari tahun 2000 hingga 2021. Thailand memiliki luas lahan pertanian paling tinggi, yaitu 43%-46%. Myanmar memiliki luas lahan pertanian paling rendah, yaitu 16%-20%.

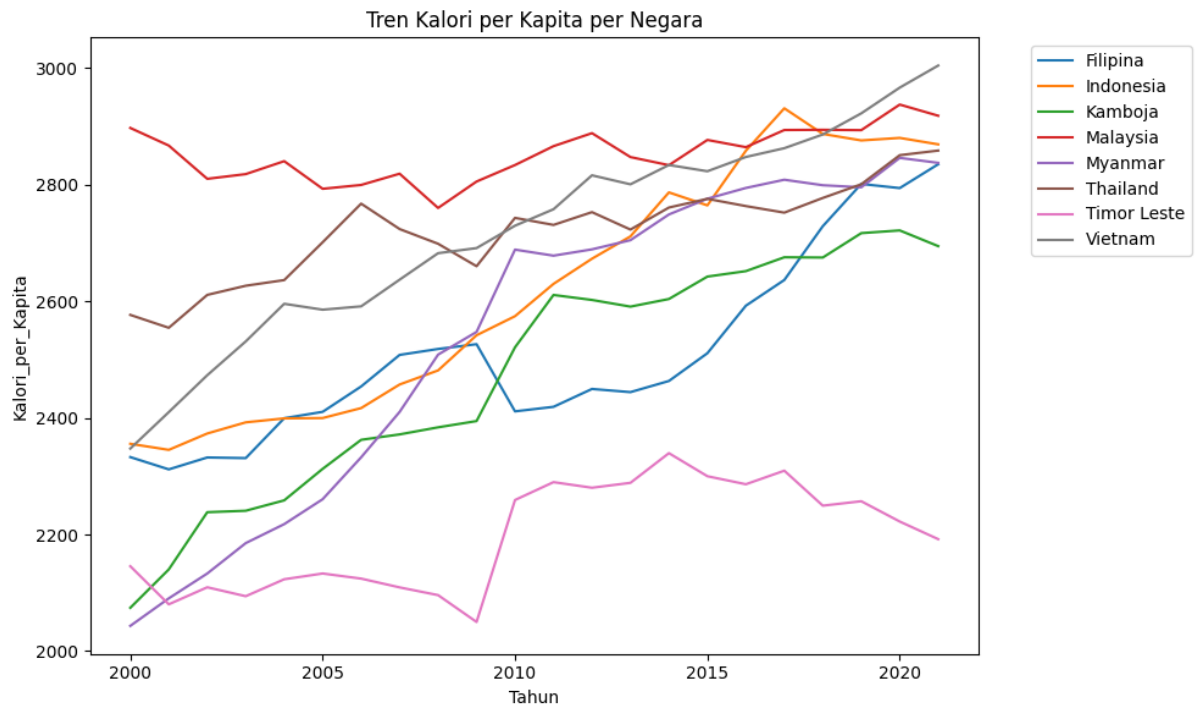
10. Visualisasi Tren Kalori per Kapita

```
plt.figure(figsize=(10,6))

sns.lineplot(
    data=DATA,
    x='Tahun',
    y='Kalori_per_Kapita',
    hue='Nama Negara'
)

plt.title('Tren Kalori per Kapita per Negara')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.savefig('tren_kalori.pdf', format='pdf')
```

Program visualisasi tren kalori per kapita dimulai dengan membuat figure berukuran 10x6. Selanjutnya, fungsi lineplot dari seaborn digunakan, dimana dataset yang digunakan adalah DATA, sumbu x adalah tahun, sumbu y adalah kalori per kapita, dan setiap negara diberi warna yang berbeda. Kemudian dengan plt.title() untuk membuat judul grafik. Selain itu, diatur juga posisi legenda grafik supaya tidak berada di area kurva, fungsi bbox_to_anchor digunakan untuk menentukan titik koordinat legenda, dan fungsi loc untuk menentukan posisi relatifnya. Setelah itu, plt.tight_layout() dipanggil untuk memastikan tata letak antar plot tidak saling bertabrakan. Terakhir, plt.savefig digunakan untuk menyimpan hasil plot dalam format pdf



Grafik di atas menunjukkan bahwa negara-negara di Asia Tenggara mengalami peningkatan konsumsi kalori per kapita sejak tahun 2000 hingga 2021. Vietnam dan Myanmar menunjukkan peningkatan paling signifikan. Malaysia dan Thailand memiliki tingkat kalori per kapita yang paling tinggi pada tahun 2000, namun trennya tidak naik secara ekstrem dan fluktuasinya tetap berada dirata-rata. Sedangkan Timor Leste memiliki kalori per kapita paling rendah sepanjang 2000-2021 dengan pola naik turun yang tidak stabil.

11. Visualisasi Tren Produksi Pangan

```
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.backends.backend_pdf import PdfPages

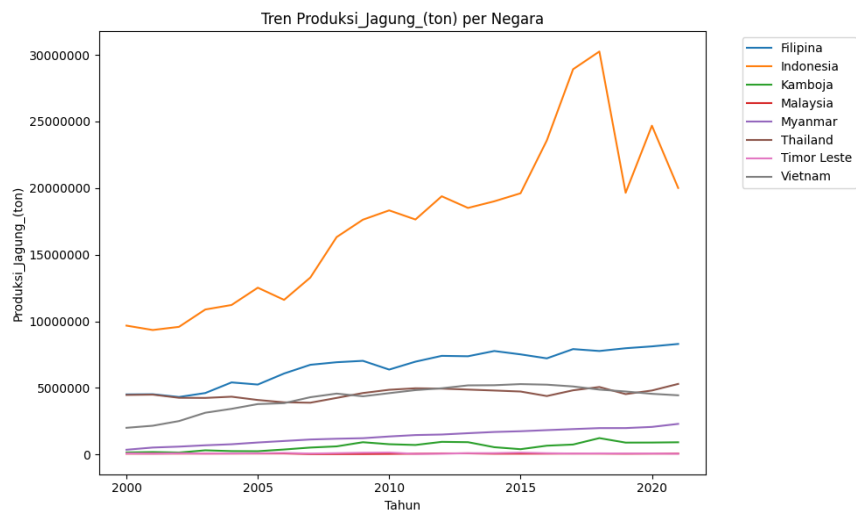
produksi_list = [
    'Produksi_Jagung_(ton)', 'Produksi_Beras_(ton)',
    'Produksi_Minyak_Sawit_(ton)', 'Produksi_Kopi_(ton)',
    'Produksi_Kakao_(ton)', 'Produksi_Teh_(ton)',
    'Produksi_Ubi_Jalar_(ton)'
]

with PdfPages('tren_produksi.pdf') as pdf:
    for var in produksi_list:
        plt.figure(figsize=(10, 6))
        sns.lineplot(
            data=DATA,
            x='Tahun',
            y=var,
            hue='Nama Negara'
        )
        plt.ticklabel_format(style='plain', axis='y')
        plt.title(f'Tren {var} per Negara')
        plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
        plt.tight_layout()

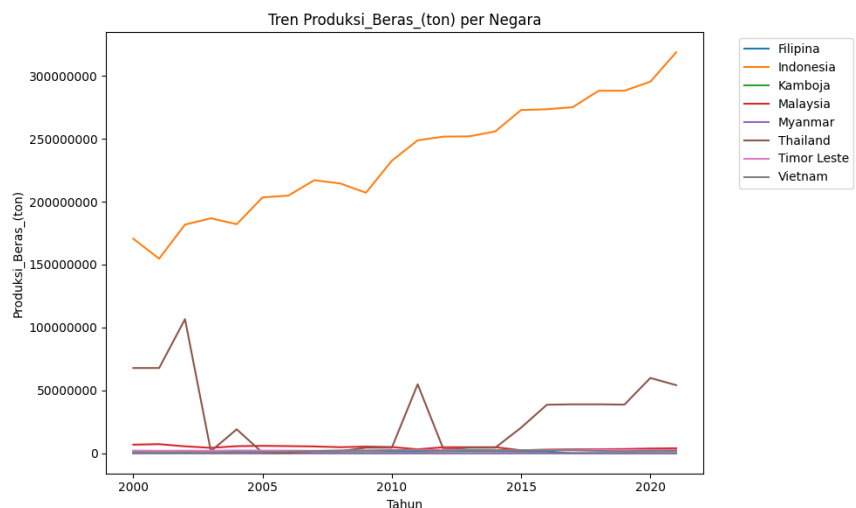
    pdf.savefig()
```

Program visualisasi tren produksi dimulai dengan mengimpor library matplotlib dan seaborn, serta modul PdfPages yang digunakan untuk menyimpan seluruh grafik dalam

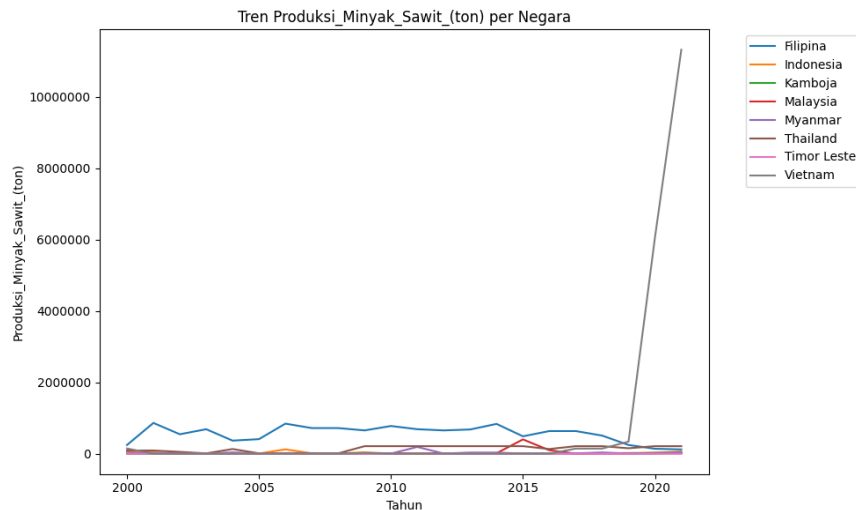
satu file PDF. Daftar variabel produksi disimpan dalam sebuah list bernama `produksi_list`, yang berisi tujuh jenis komoditas pertanian. Kemudian, blok `with PdfPages('tren_produksi.pdf')` as `pdf`: dibuat untuk membuka file PDF dan memastikan seluruh grafik yang dihasilkan akan disimpan secara otomatis. Di dalam blok tersebut, program melakukan iterasi untuk setiap variabel dalam daftar. Pada setiap iterasi, sebuah figure dengan ukuran 10x6 dibuat, lalu fungsi `sns.lineplot()` digunakan untuk menampilkan tren produksi berdasarkan tahun, dengan negara sebagai faktor pembeda melalui parameter `hue`. Fungsi `plt.ticklabel_format()` digunakan untuk menampilkan nilai sumbu y dalam format angka biasa agar lebih mudah dibaca. Setelah itu, judul grafik dibuat menggunakan `plt.title()`, dan legenda diletakkan di luar area grafik dengan bantuan parameter `bbox_to_anchor` dan `loc` agar kurva tidak tertutup. Fungsi `plt.tight_layout()` digunakan supaya tata letak grafik tetap rapi dan tidak saling bertumpukan. Terakhir, setiap grafik disimpan ke dalam file PDF menggunakan `pdf.savefig()`, sehingga seluruh grafik tren produksi dari setiap komoditas tersimpan rapi dalam 1 pdf.



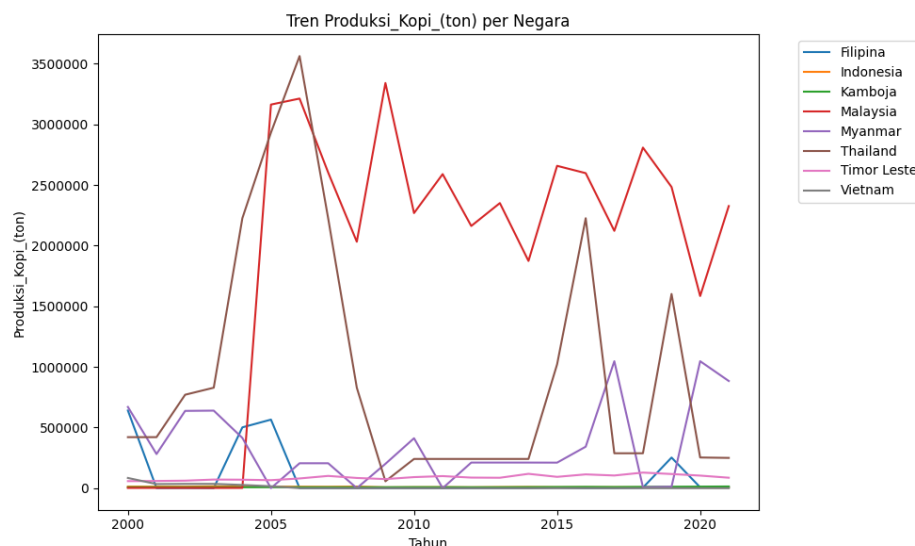
Interpretasi: Berdasarkan grafik tersebut, dapat dilihat bahwa produksi jagun di negara Indonesia mengalami tren naik dari tahun 2000-2021, sedangkan negara lain tidak mengalami tren yang signifikan, polanya cenderung horizontal dan berada di rata-ratanya.



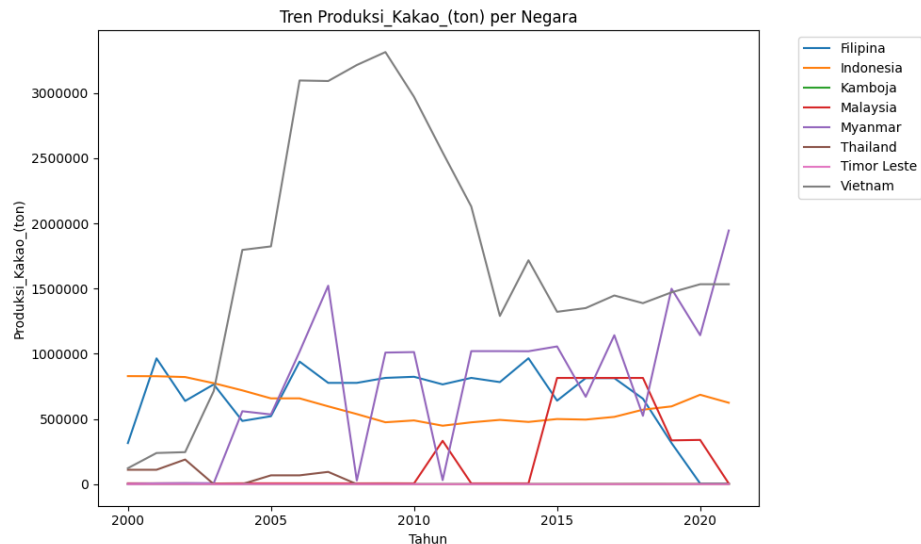
Interpretasi: Berdasarkan grafik tersebut, dapat dilihat bahwa produksi beras di negara Indonesia mengalami tren naik dari tahun 2000-2021 dan menjadi produsen beras tertinggi diantara negara lain. Negara Thailand mengalami pola fluktuatif yang tidak beraturan. Sedangkan negara lain produksi beras cenderung rendah dan tidak mengalami tren yang signifikan, polanya cenderung horizontal dan berada di rata-ratanya.



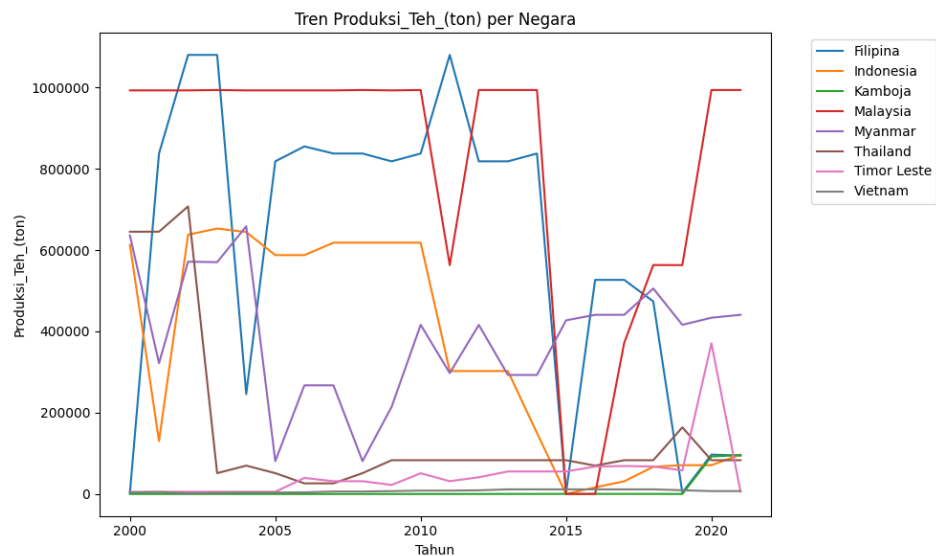
Interpretasi: Berdasarkan grafik tersebut pada tahun 2000-2019, dapat dikatakan bahwa tidak terjadi pola tren yang signifikan, produksi minyak sawit cenderung stabil. Namun, pada tahun 2019 terjadi kenaikan produksi minyak sawit secara drastis di negara Vietnam.



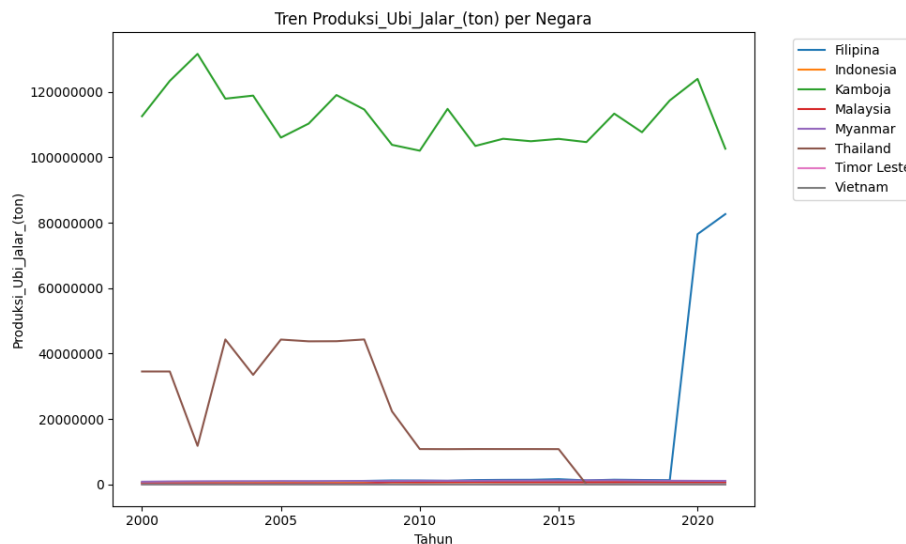
Interpretasi: Berdasarkan grafik di atas, pola produksi kopi tiap tahun cenderung fluktuatif. Namun, pada negara Malaysia dan Thailand terjadi pola fluktuatif yang sangat signifikan. Untuk negara lain cenderung stabil dan berada di rata-ratanya.



Interpretasi: Berdasarkan grafik, dapat dilihat bahwa hampir seluruh negara memiliki pola fluktuatif yang signifikan. Namun negara Vietnam memimpin dan memiliki nilai produksi kako yang tinggi. Namun, untuk negara Timor Leste nilai produksi kakaonya cenderung kecil dan polanya horizontal.



Interpretasi: Berdasarkan grafik tersebut, dapat disimpulkan bahwa hampir semua negara memiliki pola fluktuatif yang sangat signifikan dan tidak beraturan. Sehingga tidak ada tren naik atau turun yang stabil.



Interpretasi: Berdasarkan grafik tersebut, dapat dilihat bahwa negara Kamboja memiliki rata-rata nilai produksi ubi jalar tertinggi dengan pola horizontal. Negara Thailand juga memiliki pola fluktuatif namun ada pola tren turun dari tahun 2015. Kemudian negara Filipina terjadi tren naik yang sangat curam pada tahun 2019-2021. Sedangkan untuk negara-negara lain memiliki pola stabil dan horizontal dengan nilai yang cenderung kecil.

12. Transformasi Data

```
DATA_transformed = DATA[id_cols + num_cols].copy()

for c in num_cols:
    DATA_transformed[c] = np.log1p(DATA_transformed[c])

DATA_transformed.to_csv("DATA_TRANSFORMASI.csv", index=False)
DATA_transformed.head()
```

Program transformasi log dimulai dengan membuat sebuah dataset baru bernama `DATA_transformed` yang berisi `id_cols` dan `num_cols` dari dataset asli supaya proses transformasi tidak mengubah dataset awal. Setelah itu dilakukan looping untuk setiap nilai kolom pada `num_cols`. Selanjutnya diterapkan transformasi `np.log1p` dimana pada tahap ini menghitung $\log(1+x)$. Kemudian, nilai hasil transformasi `log1p` menjadi isi dari kolom baru tersebut. Kemudian dataset yang telah ditransformasi disimpan ke file CSV bernama `DATA_TRANSFORMASI.csv`. Berikut ini data hasil transformasi:

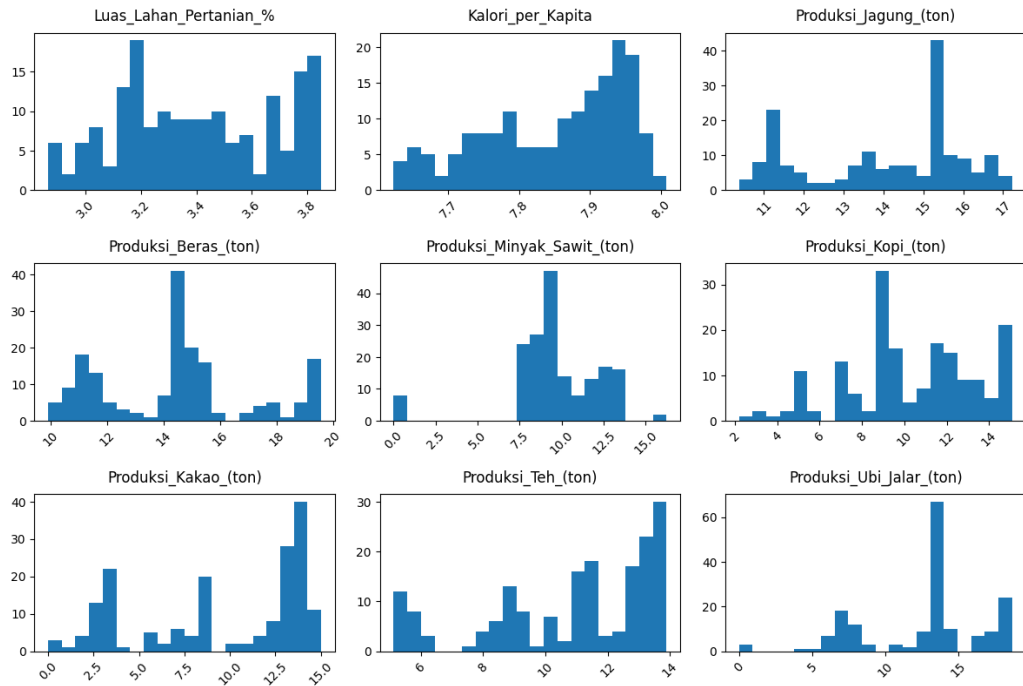
Setelah dilakukan transformasi log, perlu untuk memastikan bahwa distribusi data telah berubah.

```
axes = DATA_transformed[num_cols].hist(bins=20, figsize=(12, 8))

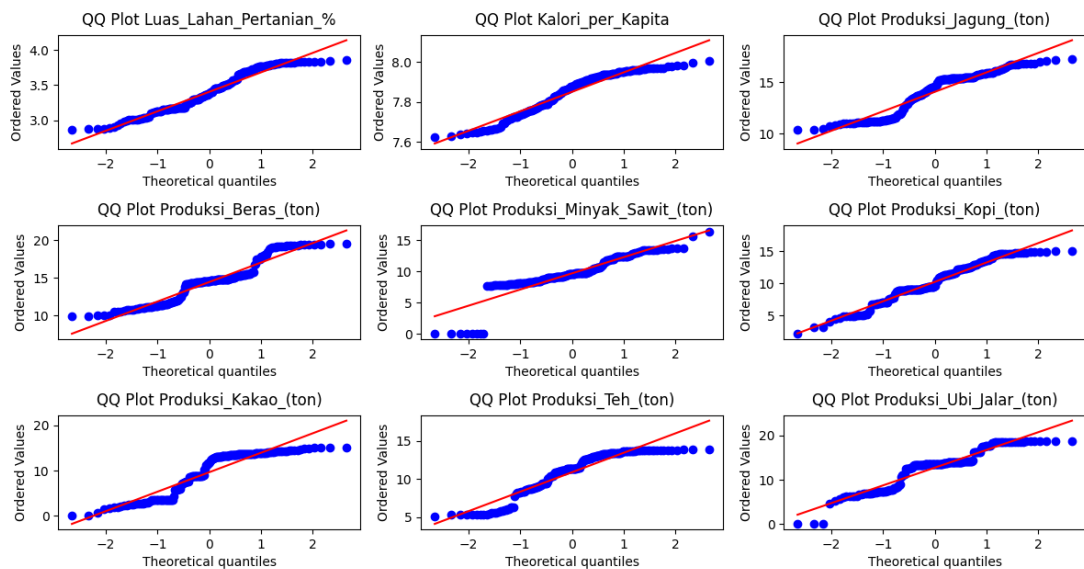
for ax in axes.flatten():
    ax.set_title(ax.get_title(), pad=10)
    ax.tick_params(axis='x', rotation=45)
    ax.grid(False)

plt.tight_layout()
plt.show()
```

Secara garis besar, program untuk membuat histogram sama dengan yang sebelumnya, hanya saja perbedaan terdapat pada variabel yang digunakan.



Berdasarkan grafik-grafik di atas, dapat dilihat bahwa distribusi data telah berubah dan tidak terlalu miring. Hal ini menunjukkan bahwa transformasi log efektif dalam mengatasi skewness dan membuat bentuk distribusi mendekati distribusi normal. Hal ini juga dibuktikan dengan Q-Q Plot berikut:

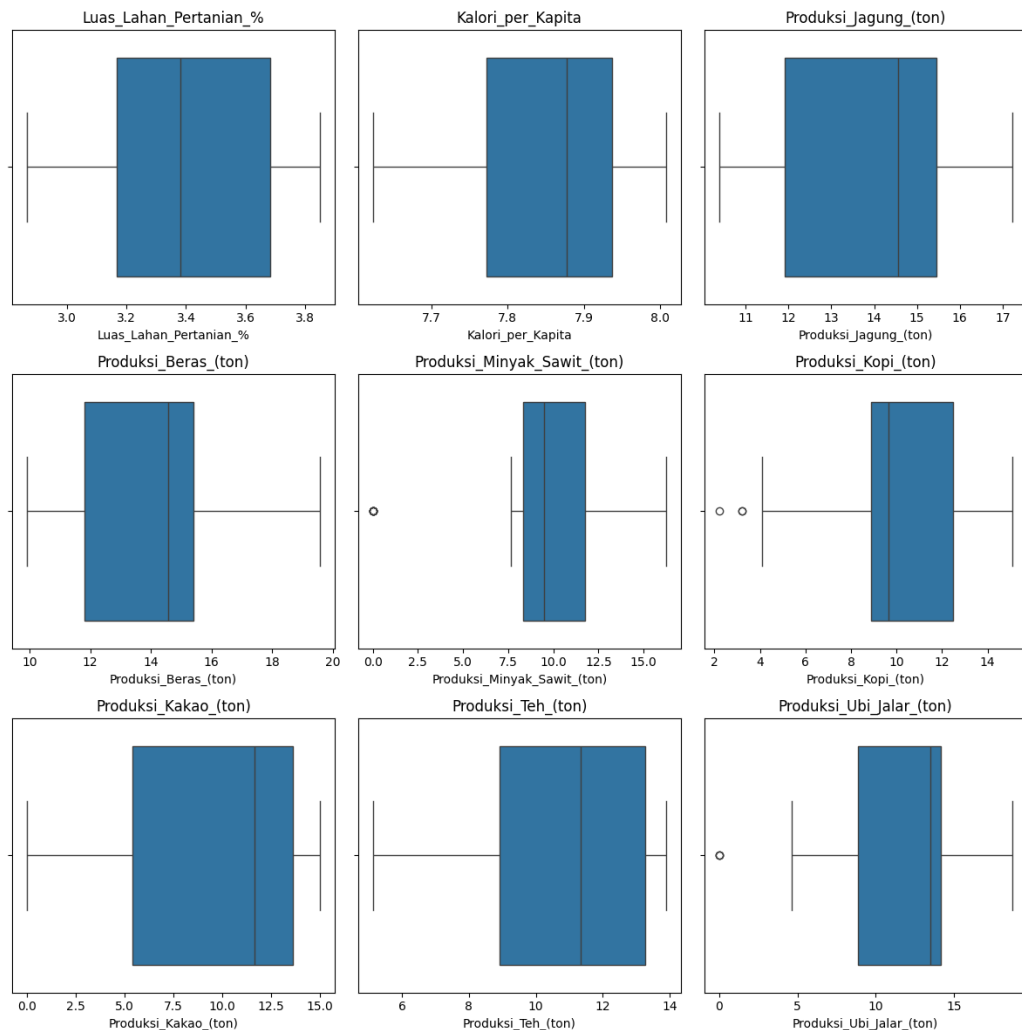


Selain pengecekan distribusi menggunakan histogram dan Q-Q Plot, dilakukan juga pengecekan outlier pada data setelah ditransformasikan.

```
plt.figure(figsize=(12, 4 * ((len(num_cols) + 2) // 3)))

for i, col in enumerate(num_cols):
    plt.subplot((len(num_cols) + 2) // 3, 3, i+1)
    sns.boxplot(x = DATA_transformed[col])
    plt.title(col)

plt.tight_layout()
plt.show()
```



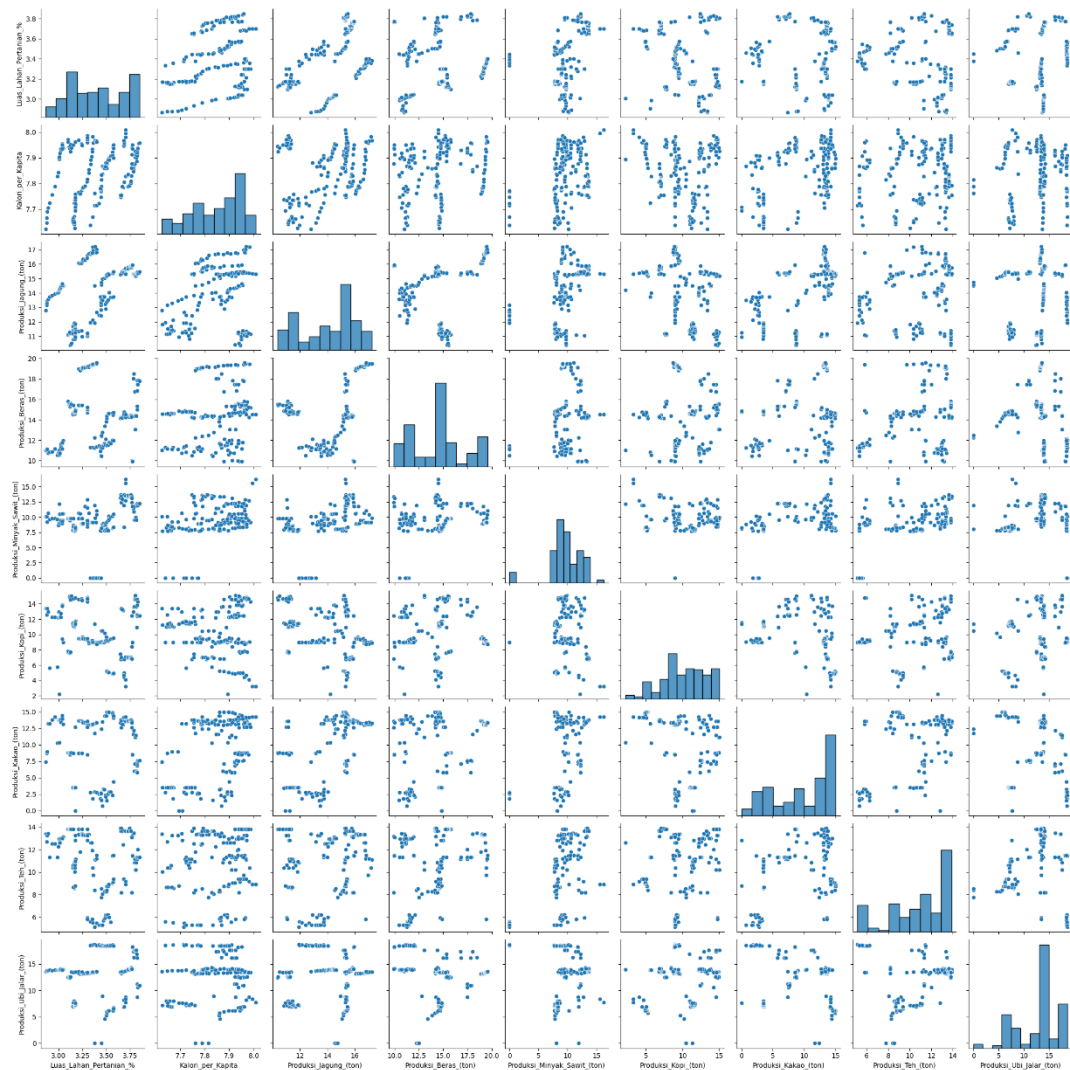
Berdasarkan box-plot sebelum dan sesudah transformasi, terlihat bahwa data yang semula memiliki banyak outlier menjadi jauh lebih terpusat. Hal ini terjadi karena transformasi log menekan nilai-nilai besar secara signifikan, sehingga perbedaan ekstrem antar nilai tidak sejauh sebelumnya. Hal ini berakibat pada sebaran data yang terlihat lebih merata dan rentangnya mengecil.

13. Analisis Korelasi

Pertama, dilakukan visualisasi hubungan data menggunakan scatterplot. Dimana menggunakan library seaborn dan matplotlib.pyplot untuk membangun visualisasi scatterplot. Dimana, dataset yang digunakan adalah kolom numerik pada DATA_transformed. Kemudian hasil plot disimpan dalam bentuk .pdf.

```
sns.pairplot(DATA[num_cols])
plt.show()
plt.savefig('scatterplot_hubungan_variabel.pdf', format='pdf')
```

Didapatkan hasil visualisasi scatterplot sebagai berikut:



Interpretasi: Scatter plot matrix tersebut memberikan gambaran awal mengenai hubungan antar variabel produksi pertanian, luas lahan, dan kalori per kapita. Secara umum, tidak terlihat pola linear yang kuat antara sebagian besar pasangan variabel, menunjukkan bahwa hubungan antar indikator cenderung lemah atau tidak berbentuk garis lurus. Secara keseluruhan, scatter plot ini mengindikasikan bahwa hubungan antar variabel masih kompleks dan kemungkinan membutuhkan analisis lanjutan seperti korelasi numerik atau pemodelan untuk memperoleh pemahaman yang lebih kuat.

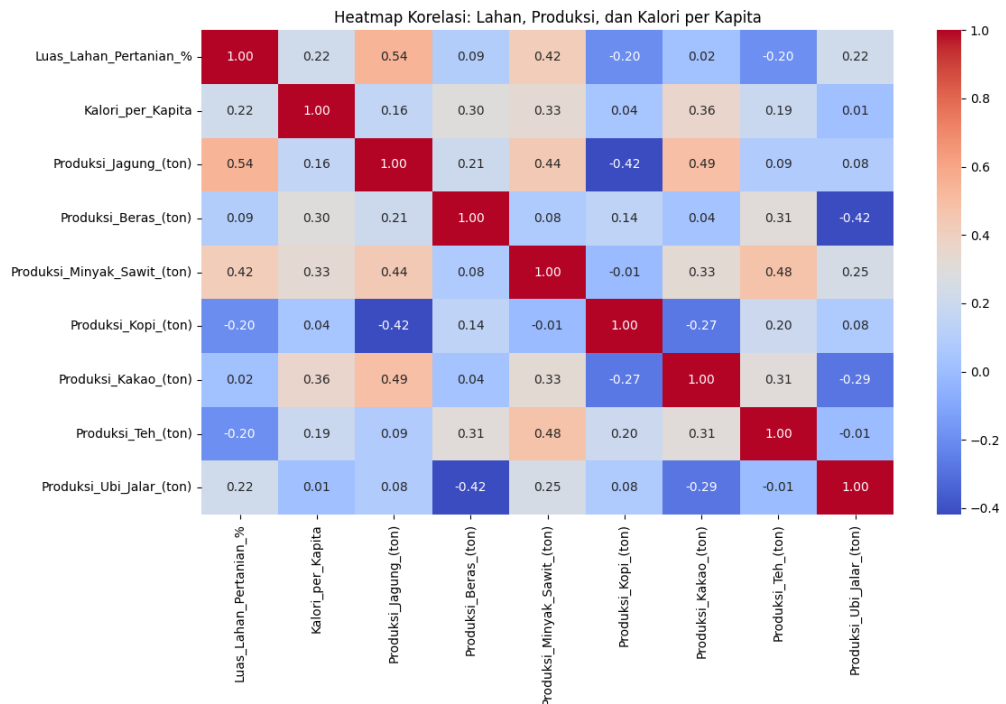
```
korelasi_spearman = DATA_transformed[num_cols].corr(method='spearman')
korelasi_spearman.to_csv("KORELASI_SPEARMAN.csv", index=False)
korelasi_spearman
```

Analisis korelasi dimulai dengan perhitungan matriks korelasi pearson menggunakan fungsi `.corr(method='spearman')` dan menggunakan dataframe `DATA_transformed` berdasarkan kolom yang sudah ditentukan di `num_cols`. Kemudian hasilnya adalah matriks korelasi yang menunjukkan hubungan linear antar variabel dengan nilai korelasi antara -1 hingga 1. Setelah itu matriks korelasi pearson yang dihasilkan akan diekspor menjadi file CSV.

Selain itu, untuk memudahkan memahami matriks korelasi, hasil tersebut dapat dibuat visualisasi dengan menggunakan heatmap.

```
plt.figure(figsize=(12,8))
sns.heatmap(korelasi_spearman, annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Heatmap Korelasi: Lahan, Produksi, dan Kalori per Kapita")
plt.tight_layout()
plt.show()
```

Program visualisasi heatmap dilakukan dengan pertama-tama membuat area plot ukuran 12x8 inch. Kemudian dengan library seaborn, heatmap dibuat menggunakan parameter seperti data matriks korelasi yang ingin divisualisasikan (korelasi_spearman), kemudian annot=True untuk menampilkan angka hasil korelasi pada masing-masing sel, cmap='coolwarm' untuk menunjukkan jika nilai korelasi negatif maka akan berwarna biru dan nilai korelasi positif akan berwarna merah, dan fmt='.2f' untuk menampilkan angka korelasi dengan format 2 angka dibelakang koma. Kemudian judul heatmap didefinisikan menggunakan fungsi plt.title(). Setelah itu, plt.tight_layout() dipanggil untuk memastikan tata letak tidak saling bertabrakan. Terakhir, plt.show() dipanggil untuk menampilkan heatmap yang telah dibuat.



Berdasarkan analisis korelasi spearman, dapat dilihat bahwa:

- Korelasi paling kuat yaitu korelasi antara Luas Lahan dengan Produksi Jagung, yaitu sebesar 0,54. Hal ini dapat disimpulkan bahwa jika luas lahan suatu negara semakin luas, maka produksi jagung juga naik.
- Korelasi paling lemah yaitu korelasi antara produksi ubi jalar dengan kalori per kapita, yaitu sebesar 0,01. Hal ini dapat disimpulkan bahwa tidak ada pengaruh yang kuat, bahkan hampir tidak ada pengaruh antara produksi ubi jalar dan kalori per kapita.

14. Standarisasi Data

```
scaler = StandardScaler()
scaled_values = scaler.fit_transform(data_mean)

data_scaled = pd.DataFrame(
    scaled_values,
    columns=data_mean.columns,
    index=data_mean.index
)
```

Pada tahap ini dilakukan proses standarisasi data menggunakan metode *StandardScaler* dari library *scikit-learn*. Standarisasi diperlukan agar setiap variabel memiliki skala yang sebanding sebelum dianalisis menggunakan algoritma K-Means. Hal ini penting karena variabel seperti total produksi, luas lahan, dan kalori per kapita memiliki satuan dan rentang nilai yang berbeda jauh.

data_scaled			
	Luas_Lahan_Pertanian_%	Kalori_per_Kapita	Total_Produksi
Nama Negara			
Filipina	1.178183	-0.325276	0.855166
Indonesia	-0.316052	0.226922	1.145978
Kamboja	0.224104	-0.522298	-1.450997
Malaysia	-0.763743	1.414409	0.374161
Myanmar	-1.571455	-0.215498	0.377980
Thailand	1.515023	0.773902	0.912345
Timor Leste	-0.867302	-2.092029	-1.483145
Vietnam	0.601242	0.739869	-0.731487

Output `data_scaled` menunjukkan hasil standarisasi tiga variabel utama Luas Lahan Pertanian, Kalori per Kapita, dan Total Produksi untuk delapan negara Asia Tenggara. Setelah distandarisasi, setiap nilai memiliki skala yang sama (mean = 0, standar deviasi = 1), sehingga angka positif menandakan nilai di atas rata-rata, sedangkan angka negatif menandakan nilai di bawah rata-rata. Misalnya, Thailand memiliki nilai tinggi pada semua variabel, sedangkan Timor Leste dan Kamboja berada jauh di bawah rata-rata. Hasil ini menjadi dasar dalam proses clustering karena semua variabel sudah berada pada skala yang sebanding.

15. Analisis Clustering

Analisis clustering dilakukan untuk mengelompokkan 8 negara Asia Tenggara berdasarkan tiga indikator utama, yaitu luas lahan pertanian(%), kalori per kapita, dan total produksi pangan. Tujuannya adalah untuk mengetahui tingkat kemiripan karakteristik ketahanan pangan antarnegara serta mengidentifikasi kelompok negara dengan kondisi yang relatif serupa.

Sebelum proses clustering dilakukan, data melalui tahap transformasi logaritma untuk mengurangi pengaruh skewness pada variabel produksi dan konsumsi, serta normalisasi min-max agar seluruh variabel berada pada skala yang sama. Normalisasi diperlukan karena K-means menggunakan jarak euclidean yang sensitif terhadap perbedaan skala data.

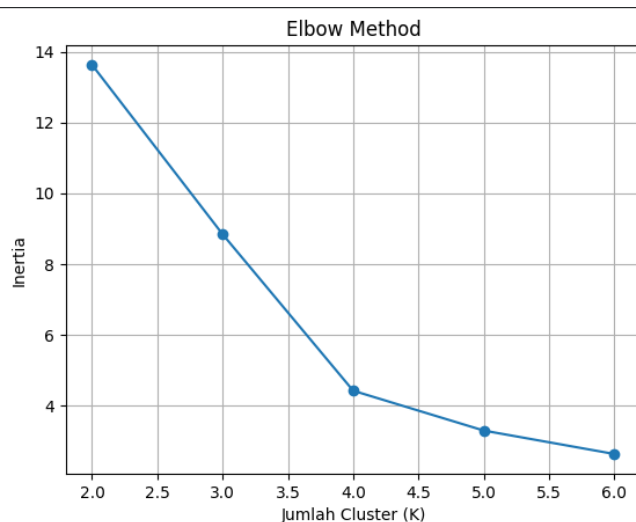
Tahap selanjutnya adalah menentukan jumlah cluster (k). Evaluasi jumlah cluster dilakukan menggunakan elbow method dan silhouette score.

```
inertia = []
K_range = range(2, 7)
X = data_scaled # pastikan numerik semua

for k in K_range:
    km = KMeans(n_clusters=k, random_state=42)
    km.fit(X)
    inertia.append(km.inertia_)

plt.plot(K_range, inertia, marker="o")
plt.title("Elbow Method")
plt.xlabel("Jumlah Cluster (K)")
plt.ylabel("Inertia")
plt.grid(True)
plt.show()
```

Tahap ini, menghitung nilai inertia pada beberapa kemungkinan jumlah cluster untuk keperluan evaluasi menggunakan metode *Elbow*. Pertama, variabel inertia dibuat sebagai list kosong untuk menyimpan hasil perhitungan inertia. Kemudian $K_range = range(2,7)$ menentukan bahwa jumlah cluster yang diuji berada pada rentang 2 hingga 6. Variabel $X = data_scaled$ memastikan bahwa data yang digunakan sudah dalam bentuk numerik dan telah melalui proses normalisasi. Selanjutnya, dilakukan perulangan `for k in K_range:` untuk menjalankan algoritma K-Means pada setiap nilai k . Di dalam perulangan tersebut, objek KMeans dibuat dengan parameter $n_clusters=k$ dan $random_state=42$ agar hasil clustering tetap konsisten. Model kemudian dilatih menggunakan data dengan `km.fit(X)` dan nilai inertia yang dihasilkan disimpan menggunakan `inertia.append(km.inertia_)`. Setelah seluruh nilai k diuji, hasil inertia divisualisasikan dalam grafik *Elbow* menggunakan `plt.plot()`. Label sumbu dan judul grafik ditambahkan melalui perintah `plt.title`, `plt.xlabel`, dan `plt.ylabel`, sementara `plt.grid(True)` mengaktifkan tampilan grid agar plot lebih mudah dibaca. Grafik ini digunakan untuk menentukan titik jumlah cluster yang paling optimal berdasarkan perubahan nilai inertia.



Gambar di atas menunjukkan hasil perhitungan Elbow Method, yaitu metode untuk menentukan jumlah cluster yang ideal dalam algoritma K-Means. Sumbu X menunjukkan jumlah cluster (K) yang diuji, sedangkan sumbu Y menunjukkan nilai inertia, yaitu total jarak data terhadap centroid dalam cluster. Terlihat bahwa penurunan inertia terbesar terjadi antara K=2 hingga K=4, kemudian grafik mulai melandai setelah K=4. Titik perubahan atau "elbow" berada di antara K=3 dan K=4, sehingga K=3 dipilih sebagai nilai yang paling sesuai karena mampu memberikan pemisahan cluster yang baik tanpa membuat model terlalu kompleks. Pilihan ini juga selaras dengan kebutuhan penelitian yang mengklasifikasikan negara ke dalam tiga kategori tingkat ketahanan pangan.

```
for k in K_range:
    km = KMeans(n_clusters=k, random_state=42)
    labels = km.fit_predict(X)
    score = silhouette_score(X, labels)
    print(f"K={k}, Silhouette Score={score}")

... K=2, Silhouette Score=0.2536060471920948
    K=3, Silhouette Score=0.2934451405671265
    K=4, Silhouette Score=0.28232256820524465
    K=5, Silhouette Score=0.17884922577368065
    K=6, Silhouette Score=0.06734960140356788
```

Selanjutnya untuk menentukan jumlah kluster yang optimal pada algoritma K-Means clustering dengan menggunakan Silhouette Score sebagai metrik evaluasi. Prosesnya dilakukan dalam sebuah perulangan (for) yang mengeksplorasi berbagai nilai k yang disimpan dalam K_range, yaitu rentang jumlah kluster yang ingin dicoba. Pada setiap iterasi, objek KMeans dibuat dengan jumlah kluster k dan random_state=42 untuk memastikan hasil yang konsisten. Kemudian, metode fit_predict(X) dijalankan pada data X, yang melakukan pelatihan model K-Means dan sekaligus menghasilkan label kluster untuk setiap data. Setelah itu, Silhouette Score dihitung menggunakan fungsi silhouette_score(X, labels), yang memberikan ukuran seberapa baik setiap objek diklusterkan: nilai mendekati 1 menunjukkan kluster yang jelas dan terpisah, sedangkan nilai mendekati -1 menunjukkan kluster yang tumpang tindih atau salah pengelompokan. Akhirnya, hasil berupa jumlah kluster k dan skor Silhouette yang sesuai dicetak menggunakan print, sehingga dapat dibandingkan untuk memilih jumlah kluster yang paling optimal.

Pemilihan jumlah cluster dievaluasi menggunakan Elbow Method dan Silhouette Score. Hasil uji menunjukkan bahwa inertia menurun tajam hingga K = 3, kemudian perlambatan terjadi setelahnya. Sementara itu, nilai silhouette tertinggi diperoleh pada K = 3 (0.293), yang menunjukkan keseimbangan antara pemisahan cluster dan jumlah data yang relatif kecil. Berdasarkan hasil tersebut, serta pertimbangan interpretasi kategorisasi ketahanan pangan (tinggi, sedang, rendah), maka jumlah cluster akhir ditetapkan sebanyak tiga cluster.

K-MEANS

```
▶ kfinal = 3
kmeans = KMeans(n_clusters=kfinal, random_state=42)
labels = kmeans.fit_predict(X)
data_scaled["Cluster"] = labels
```

Langkah ini melakukan klastering akhir pada data dengan jumlah klaster yang sudah ditentukan, yaitu tiga. Pertama, dibuat objek K-Means dengan jumlah klaster tersebut dan `random_state` agar hasilnya konsisten. Selanjutnya, model dilatih pada data X dan menghasilkan label klaster untuk setiap titik data. Label ini kemudian ditambahkan ke dataset asli sebagai kolom baru bernama "Cluster",

```
▶ cluster_mean = data_scaled.groupby("Cluster")[[
    "Luas_Lahan_Pertanian_%",
    "Kalori_per_Kapita",
    "Total_Produksi"
]].mean()

cluster_mean["Total"] = cluster_mean.sum(axis=1)
cluster_mean = cluster_mean.sort_values("Total")

urutan = cluster_mean.index.tolist()
mapping = {}
1: "Tinggi",
2: "Sedang",
0: "Rendah"]

# Terapkan mapping
data_scaled["Kategori"] = data_scaled["Cluster"].map(mapping)

data_scaled["Kategori"] = data_scaled["Cluster"].map(mapping)
```

Selanjutnya untuk memberikan label kategori pada hasil clustering berdasarkan tingkat kemiripan karakter data. Pertama, dilakukan perhitungan nilai rata-rata untuk tiga variabel utama, yaitu luas lahan pertanian, kalori per kapita, dan total produksi pada setiap cluster. Setelah itu, hasil rata-rata tersebut dijumlahkan untuk memperoleh nilai total yang digunakan sebagai dasar peringkat cluster. Cluster kemudian diurutkan dari nilai total tertinggi hingga terendah untuk menentukan kategori, yaitu Tinggi, Sedang, dan Rendah. Selanjutnya, proses pemetaan dilakukan dengan memberikan label kategori pada masing-masing anggota cluster sesuai hasil urutan tersebut. Dengan demikian, setiap negara dalam dataset tidak hanya memiliki nomor cluster, tetapi juga label yang lebih bermakna sehingga interpretasi hasil analisis menjadi lebih mudah dan informatif.

HASIL CLUSTERING MENGGUNAKAN K-MEANS

```
▶ hasil = data_scaled.copy()
hasil.to_csv("HASIL_CLUSTER_FINAL.csv")

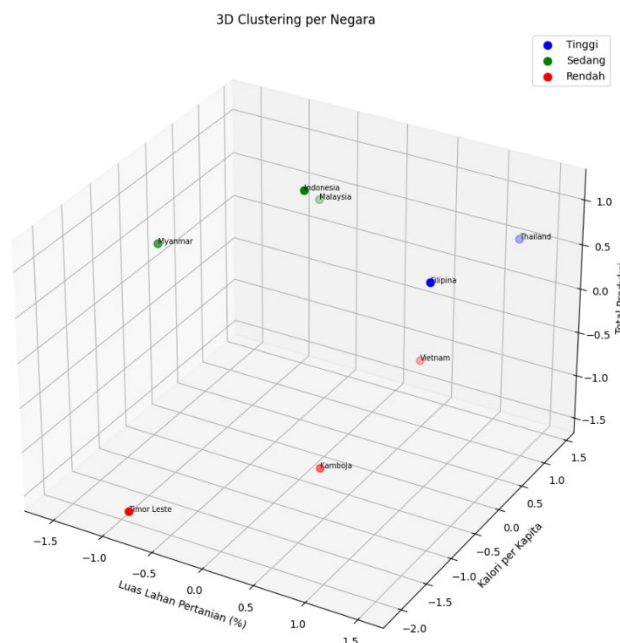
print("\nHASIL CLUSTER PER NEGERA")
print(hasil[["Cluster", "Kategori"]])
```

```
...
HASIL CLUSTER PER NEGERA
Cluster Kategori
Nama Negara
Filipina      1  Tinggi
Indonesia    2  Sedang
Kamboja       0  Rendah
Malaysia      2  Sedang
Myanmar       2  Sedang
Thailand      1  Tinggi
Timor Leste   0  Rendah
Vietnam       0  Rendah
```

Model K-Means menghasilkan tiga kategori cluster yaitu Tinggi, Sedang, dan Rendah. Penentuan label dilakukan dengan menghitung rata-rata indikator utama (Total Produksi, Luas Lahan Pertanian, Kalori per Kapita).

Negara	Kategori
Filipina	Tinggi
Indonesia	Sedang
Kamboja	Rendah
Malaysia	Sedang
Myanmar	Sedang
Thailand	Tinggi
Timor Leste	Rendah
Vietnam	Rendah

Hasil ini menunjukkan bahwa Thailand dan Filipina berada pada kategori ketahanan pangan tinggi, ditandai oleh tingkat produksi dan ketersediaan kalori yang relatif lebih baik dibanding negara lain. Indonesia, Malaysia, dan Myanmar berada pada kategori sedang, yang berarti produksi dan kalori per kapita cukup baik namun belum optimal. Sementara itu, Vietnam, Kamboja, dan Timor Leste berada pada kategori rendah, mencerminkan kapasitas produksi pangan dan akses kalori per kapita yang masih terbatas. Gambar berikut merupakan representasi clustering dalam bentuk *scatter plot 3D* berdasarkan tiga variabel utama yang dianalisis.



Visualisasi tersebut memperlihatkan bahwa cluster memiliki pemisahan yang cukup jelas, meskipun terdapat kedekatan antar-negara dalam cluster menengah. Hasil ini didukung nilai silhouette score sebesar 0.293, yang menunjukkan bahwa struktur cluster cukup stabil dan sesuai untuk dataset dengan ukuran kecil dan variabel multiskala seperti ini

4.5 Kendala dan Rencana

Aspek	Uraian
Kendala	<ul style="list-style-type: none"> - Adanya missing value, perbedaan format kolom, dan ketidakkonsistenan persamaan negara antar dataset sehingga proses cleaning memakan waktu lebih lama - Cakupan data antar 8 negara Asia Tenggara tidak merata pada beberapa tahun, sehingga dapat memengaruhi hasil analisis deskriptif dan tren - Struktur 3 data dataset berbeda-beda sehingga ketika integrasi perlu penyesuaian manual untuk menyamakan format - Data yang tidak memiliki pola linear membuat hasil analisis korelasi kurang menggambarkan hubungan sebenarnya antar variabel. - Proses eksplorasi data beberapa kali mengalami error, terutama pada tahap clustering dan plotting, akibat format data yang belum sepenuhnya konsisten.
Rencana analisis lanjutan	<p>Rencana analisis lanjutan dalam penelitian ini adalah melakukan peramalan data deret waktu menggunakan model ARIMA. Metode tersebut dipilih karena seluruh variabel penelitian, yaitu luas lahan pertanian, kalori per kapita, dan berbagai produksi komoditas pangan berbentuk data tahunan yang berpotensi memuat pola tren maupun musiman. Tahapan analisis meliputi pemeriksaan stasioneritas, penentuan ordo model, estimasi parameter, serta penilaian kinerja model menggunakan indikator seperti AIC dan RMSE untuk memperoleh model yang paling optimal. Hasil prediksi dari ARIMA diharapkan dapat memberikan gambaran mengenai arah perkembangan produksi dan pangan di masa depan, sehingga dapat mendukung proses perencanaan dan pengambilan keputusan pada sektor pertanian dan pangan.</p>

4.6 Kontribusi Kelompok

No	Tahapan Proses	Deskripsi Tugas	Penanggung Jawab
1.	Pengambilan Data, Cleaning dan Integrasi	<ul style="list-style-type: none"> - Mengambil dataset dari 3 sumber (AGR, CAL, WFP) - Filter 8 negara Asia Tenggara, 7 hasil produksi pangan - Pengecekan dan penanganan missing Value - Perbaikan format kolom dan penyamaan nama negara - Integrasi 3 dataset 	Cintiya Agustin Nareswari
2.	Transformasi	<ul style="list-style-type: none"> - Transformasi log 	Ayda Syifa Ul Aliyah
3.	Analisis Data Eksploratif	<ul style="list-style-type: none"> - Statistik deskriptif - Barplot - Histogram 	Ayda Syifa Ul Aliyah

		<ul style="list-style-type: none"> - Boxplot - Q-Q Plot - Lineplot - Density Plot - Grouping - Pivoting 	
4.	Analisis Korelasi	<ul style="list-style-type: none"> - Analisis korelasi spearman - Visualisasi hasil korelasi dengan heatmap - Visualisasi sebaran hubungan data dengan scatterplot 	Ayda Syifa Ul Aliyah
5.	Analisis Clustering	<ul style="list-style-type: none"> - Menjalankan K-Means - Menentukan jumlah cluster - Interpretasi hasil Cluster 	Cintiya Agustin Nareswari
6.	Hasil Clustering	<ul style="list-style-type: none"> - Visualisasi hasil clustering dengan scatter plot 	Ayda Syifa Ul Aliyah

BAB V

PENUTUP

5.1 Kesimpulan

Penelitian ini menunjukkan bahwa perkembangan luas lahan pertanian, produksi pangan, dan kalori per kapita di delapan negara Asia Tenggara selama 2000–2021 memiliki pola yang sangat beragam. Thailand dan Filipina memiliki proporsi lahan pertanian yang besar, sementara Myanmar dan Malaysia relatif kecil. Produksi pangan memperlihatkan perbedaan ekstrem antarnegara, di mana Indonesia unggul pada beras dan jagung, Malaysia pada minyak sawit dan kopi, serta Kamboja pada produksi ubi jalar. Konsumsi kalori per kapita meningkat di sebagian besar negara, namun tetap rendah dan tidak stabil di Timor Leste. Hasil eksplorasi awal juga menunjukkan bahwa sebagian besar variabel memiliki distribusi tidak normal sehingga dilakukan transformasi logaritma untuk memperbaiki skewness dan mengurangi efek outlier. Analisis korelasi Spearman menunjukkan hubungan yang cenderung lemah hingga sedang antarvariabel, dengan korelasi terkuat muncul pada hubungan luas lahan dan produksi jagung, serta korelasi paling lemah antara produksi ubi jalar dan kalori per kapita.

Hasil analisis clustering setelah proses transformasi dan normalisasi menghasilkan tiga kelompok negara berdasarkan karakteristik ketahanan pangannya, yaitu cluster tinggi (Filipina dan Thailand), cluster sedang (Indonesia, Malaysia, dan Myanmar), serta cluster rendah (Vietnam, Kamboja, dan Timor Leste). Cluster tinggi dicirikan oleh kapasitas produksi dan tingkat kalori yang relatif lebih tinggi, sementara cluster rendah menunjukkan keterbatasan baik pada produksi maupun konsumsi energi masyarakat. Dengan nilai silhouette score sebesar 0.293, struktur cluster dinilai cukup baik untuk dataset berskala kecil. Secara keseluruhan, penelitian ini menyimpulkan bahwa ketahanan pangan negara-negara Asia Tenggara merupakan hasil interaksi kompleks antara ketersediaan lahan, pola produksi komoditas utama, serta tingkat konsumsi masyarakat. Keragaman kondisi antarnegara menegaskan pentingnya pengambilan kebijakan pangan yang komprehensif, berbasis data, dan mempertimbangkan karakteristik khusus tiap negara.

BAB IV

DAFTAR PUSTAKA

- Heckert, N. & Filliben, J. (2003). *Chapter 1: Exploratory Data Analysis*. In *NIST/SEMATECH e-Handbook of Statistical Methods*. National Institute of Standards and Technology.
- Hakimah, Y., Ratar, M., Ardan, M., & Setiawan, A. (2024). Analisis data statistik: Panduan komprehensif untuk interpretasi data. PT Media Penerbit Indonesia.
- Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioral sciences* (10th ed.). Cengage Learning.
- McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.
- Wickham, H. (2014). *Tidy Data*. *Journal of Statistical Software*, 59(10), 1–23.
- Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. *Pattern Recognition Letters*, 31(8), 651–666.
- Pratama, F. (2019). *Data Mining: Teknik dan Implementasi Algoritma Clustering untuk Pengelompokan Data*. Informatika.
- Rahm, E., & Do, H. H. (2000). *Data Cleaning: Problems and Current Approaches*. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

Lampiran 1. Sumber data, kode program, hasil proses wrangling, file presentasi, file laporan, serta seluruh berkas pendukung projek dapat diakses melalui repository GitHub pada tautan berikut:

https://github.com/Cintiyanareswari/Kelompok5_Cintiya_dan_Ayda_Projek_Data_Wrangling