LAPORAN TUGAS AKHIR DATA WRANGLING: ANALISIS HUBUNGAN ANTARA UPAH MINIMUM PROVINSI, POLA MAKAN, DAN PREVALENSI KESEHATAN DI INDONESIA

Penyusun:

Salsabila Alika Seftizianka (24031554117)

Naufal Muzaki (24031554061)

Kelas 2024 D

Mata Kuliah:

Data Wrangling

Dosen Pengampu:

Dinda Galuh Guminta, S.Stat., M.Stat.

Belgis Ainatul Iza, S.Si., M.Mat.

Program Studi S1 Sains Data

Fakultas Matematika Dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

DAFTAR ISI

BAB I

PENDAHULUAN

A. Latar Belakang

Indonesia merupakan negara dengan kondisi ekonomi, sosial dan budaya yang sangat beragam. Keragaman ini menciptakan dinamika yang kompleks dalam pemenuhan kebutuhan dasar, termasuk pangan dan kesehatan. Salah satu indikator penting untuk melihat kondisi ekonomi suatu wilayah adalah Upah Minimum Provinsi (UMP). Variasi UMP mempengaruhi daya beli masyarakat dalam memenuhi kebutuhan hidup termasuk kemampuan untuk memperoleh makanan yang bergizi. Namun tingginya UMP tidak selalu menjamin kualitas konsumsi pangan yang baik, karena faktor harga pangan, preferensi budaya, dan akses terhadap makanan sehat turut berperan.

Di sisi lain, pola makan masyarakat Indonesia masih menghadapi sejumlah tantangan. Seperti tingginya ketergantungan pada makanan pokok, rendahnya konsumsi buah dan sayur, serta meningkatnya konsumsi makanan instan dan minuman manis menunjukkan bahwa pilihan makanan seringkali ditentukan oleh faktor ekonomi dan gaya hidup modern. Masyarakat berpendapatan lebih tinggi memiliki akses lebih luas terhadap makanan yang bernutrisi sedangkan kelompok berpendapatan rendah cenderung memilih makanan yang murah tetapi rendah nutrisi.

Kondisi ini berkaitan erat dengan meningkatnya masalah kesehatan masyarakat, termasuk stunting, obesitas, anemia, dan penyakit tidak menular. Fenomena beban gizi ganda memperlihatkan bahwa Indonesia menghadapi dua jenis masalah yaitu, kekurangan gizi di beberapa wilayah dan kelebihan gizi di wilayah lainnya

Melihat fenomena yang terjadi di masyarakat , perlu untuk dilakukan analisis berbasis data untuk mengetahui apakah terdapat hubungan antara UMP, pola makan, dan prevalensi kesehatan di Indonesia. data dari berbagai sumber dapat digabungkan sehingga memungkinkan analisis yang lebih baik.

B. Rumusan Masalah

1. Bagaimana mengintegrasikan data UMP, asupan kalori, dan kesehatan dari berbagai sumber menggunakan teknik data wrangling?
2. Apakah terdapat hubungan antara tingkat UMP dan pola asupan kalori masyarakat di Indonesia?
3. Bagaimana pengaruh pola asupan kalori terhadap prevalensi kesehatan

C. Tujuan

1. Mengumpulkan dan mengolah data UMP, asupan kalori, dan kesehatan dari sumber yang diperoleh.
2. Menganalisis distribusi UMP, asupan kalori, dan indikator kesehatan antarprovinsi.
3. Mengidentifikasi hubungan antara UMP dan pola asupan kalori masyarakat.
4. Menganalisis pengaruh pola asupan kalori terhadap prevalensi kesehatan berdasarkan dataset yang sudah digabungkan.
5. Menyediakan visualisasi data yang telah diolah.

D. Manfaat

1. Manfaat bagi Akademisi: Penelitian ini memberikan kontribusi dalam pengembangan literatur terkait hubungan antara faktor ekonomi, pola makan dan kesehatan. Penelitian ini juga sebagai bentuk contoh nyata penerapan teknik data wrangling dalam analisis multidisipliner.
2. Manfaat bagi Pemerintah: Hasil penelitian dapat menjadi bahan pertimbangan dalam penyusunan kebijakan upah minimum, kebijakan ketahanan pangan, serta program peningkatan kesehatan masyarakat.
3. Manfaat bagi Masyarakat: Penelitian ini memberikan wawasan mengenai pentingnya pola makan sehat dan bagaimana faktor ekonomi mempengaruhi konsumsi pangan. Masyarakat dapat memahami pentingnya pemilihan makanan bergizi meskipun dalam keterbatasan ekonomi.
4. Manfaat bagi Penulis: Penulis mendapatkan pengalaman langsung dalam menerapkan teknik data wrangling seperti scraping, ekstraksi PDF, pembersihan data, analisis statistik, dan visualisasi. Selain itu penelitian ini meningkatkan kemampuan penulis dalam menyusun laporan dan melakukan analisis data.

TINJAUAN PUSTAKA

A. Upah Minimum Provinsi

Upah Minimum Provinsi (UMP) merupakan salah satu indikator penting yang digunakan untuk menilai kondisi ekonomi suatu wilayah. UMP berfungsi sebagai jaring pengaman sosial bagi pekerja agar memperoleh penghasilan yang layak sehingga dapat memenuhi kebutuhan dasar, termasuk pangan, kesehatan, dan pendidikan. Variasi UMP antar provinsi dipengaruhi oleh pertumbuhan ekonomi, biaya hidup, inflasi, kebutuhan hidup layak, serta struktur industri di masing-masing daerah. Penelitian Durrotun (2024) menunjukkan bahwa UMP memiliki pengaruh signifikan terhadap kesejahteraan masyarakat Indonesia, terutama melalui peningkatan daya beli dan kemampuan rumah tangga dalam memenuhi kebutuhan dasar. Oleh karena itu, variasi UMP antarprovinsi dapat berdampak langsung terhadap daya beli masyarakat dan pola konsumsi rumah tangga.

B. Pola Konsumsi Energi, Protein, dan Gizi Masyarakat Indonesia

Pola makan masyarakat Indonesia masih menghadapi tantangan dalam pemenuhan gizi seimbang. Laporan resmi Badan Pusat Statistik (BPS) mengenai konsumsi energi dan protein penduduk Indonesia menunjukan sebagian besar masyarakat masih cenderung bergantung pada konsumsi karbohidrat dengan variasi tingkat konsumsi energi dan protein antar provinsi (BPS, 2024). Distribusi pemenuhan gizi tersebut dipengaruhi oleh faktor ekonomi, sosial, budaya, dan ketersediaan pangan di suatu wilayah. Data kuintil konsumsi energi dan protein dari BPS juga menunjukkan adanya ketimpangan konsumsi antar kelompok pendapatan, terutama antara kuintil pertama dan kelima sehingga diperlukan analisis untuk melihat keterkaitannya dengan kondisi ekonomi masing-masing provinsi.

C. Prevalensi Kesehatan: Gizi Buruk, Obesitas, dan Penyakit Terkait Pola Makan

Kondisi kesehatan masyarakat terutama terkait gizi merupakan indikator penting dalam pembangunan manusia. Berdasarkan laporan UNICEF mengenai lanskap obesitas dan kelebihan berat badan di Indonesia, pola konsumsi makanan modern seperti makanan olahan, minuman manis, serta rendahnya konsumsi sayur dan buah berkontribusi terhadap meningkatnya prevalensi obesitas di berbagai provinsi (UNICEF, 2024). Sementara itu, stunting banyak ditemukan pada wilayah yang memiliki keterbatasan akses pangan bergizi dan konsumsi protein hewani yang rendah. Hal ini menunjukkan adanya hubungan yang kuat antara faktor ekonomi, pola makan, dan kondisi kesehatan. Oleh karena itu analisis diperlukan agar dapat memahami penyebab ketimpangan kesehatan secara lebih lengkap.

D. Data Wrangling sebagai Metode Integrasi dan Analisis Data

Data wrangling merupakan proses penting dalam analisis data modern yang mencakup pengumpulan, pembersihan, transformasi dan integrasi data dari berbagai sumber. Dalam penelitian ini data wrangling diperlukan karena setiap dataset seperti data UMP ,data konsumsi energi dan protein dari publikasi BPS, serta data kesehatan dari laporan Unicef memiliki struktur, rentang nilai dan kualitas data yang berbeda. Oleh karena itu, penerapan data wrangling menghasilkan integrasi data yang lebih baik dan akurat sehingga analisis hubungan antara UMP, pola makan, dan kesehatan dapat dilakukan dengan optimal.

E. Penelitian Terdahulu

Beberapa penelitian sebelumnya telah menunjukkan bahwa pendapatan keluarga memiliki hubungan erat dengan pola konsumsi pangan dan status gizi. Hasan (2015) menemukan bahwa pendapatan yang lebih tinggi berdampak pada peningkatan konsumsi energi dan protein, sehingga status gizi pekerja menjadi lebih baik. Temuan ini menegaskan bahwa kemampuan ekonomi mempengaruhi kualitas konsumsi pangan harian.

Penelitian serupa dilakukan oleh Ningsih dan Masrikhiyah (2021), yang mengkaji pengaruh pendapatan dan kecukupan energi terhadap status gizi ibu hamil. Mereka menyimpulkan bahwa pendapatan berpengaruh signifikan terhadap kecukupan energi, yang kemudian berkaitan dengan risiko Kurang Energi Kronis (KEK). Hal ini menunjukkan bahwa kondisi ekonomi berperan penting dalam pemenuhan kebutuhan gizi kelompok rentan.

Selain itu, Naibaho dan Aritonang (2021) meneliti hubungan pendapatan dan pengetahuan gizi ibu terhadap ketahanan pangan keluarga. Studi tersebut menemukan bahwa pendapatan merupakan faktor dominan yang mempengaruhi kemampuan keluarga dalam menyediakan pangan bergizi. Pendapatan yang rendah meningkatkan kemungkinan keluarga mengalami keterbatasan akses pangan berkualitas.

Secara keseluruhan penelitian-penelitian ini menunjukkan bahwa pendapatan yang dalam konteks penelitian ini direpresentasikan melalui Upah Minimum Provinsi (UMP) berkaitan dengan pola konsumsi energi dan protein serta berdampak pada kondisi gizi dan kesehatan masyarakat. Oleh karena itu, penelitian ini relevan untuk melihat hubungan antara UMP, pola makan dan prevalensi kesehatan antar provinsi di Indonesia.

BAB III

METODOLOGI

A. Sumber Data

1. Data UMP 2024 dari situs Catapa.
   Pada website Catapa berisikan upah minimum tiap provinsi secara lengkap. Provinsi pada website ini berjumlah 38 provinsi.

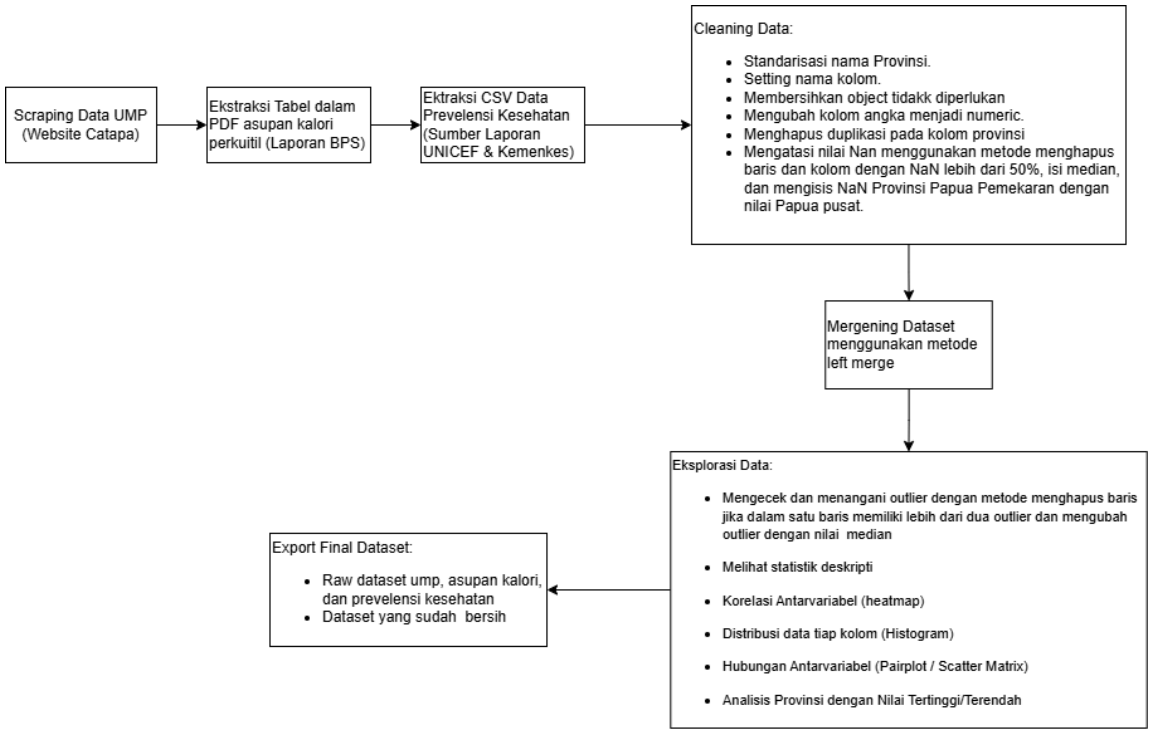
2. Data Asupan Kalori dari PDF resmi BPS.
   Data asupan kalori berasal dari laporan resmi BPS. Pada laporan ini terdapat data asupan kalori tiap provinsi berdasarkan kuintil. Kuintil merupakan kelompok kelas ekonomi. JIka kuintil 1, maka data tersebut adalah jumlah asupan kalori berdasarkan kelas ekonomi terbawah (termiskin).

3. Data kesehatan masyarakat dari PDF laporan UNICEF.
   Data ini bersumber dari PDF laporan UNICEF dan Kemenkes yang berisikan prevalensi kesehatan. Prevalensi kesehatan yaitu berupa data stunting, wasting, kelebihan berat badan secara umum, kelebihan berat badan berdasarkan kategori umur, dan kerawanan pangan. Prevalensi ini disajikan dalam bentuk per-provinsi.

B. Variabel yang Digunakan

1. Variabel Ekonomi: UMP.
2. Variabel Asupan Kalori: Rata-rata asupan kalori setiap provinsi yang diklasifikasikan berdasarkan kelas ekonomi (kuintil pertama–kuintil kelima)
3. Variabel Kesehatan: Stunting, wasting, kelebihan berat badan secara umum, kelebihan berat badan berdasarkan kategori umur, dan kerawanan pangan.

C. Pipeline



D. Teknik Pengambilan Data dan Integrasi Data

1. Mengambil data UMP dari Situs web Catapa menggunakan scraping.

```python
import requests
from bs4 import BeautifulSoup
import pandas as pd
import re


def rupiah_to_int(s):
    s = s.strip()
    s = s.replace("Rp", "").replace(".", "").replace(",", "")
    try:
        return int(s)
    except:
        return None

def norm_prov(prov):
    prov = prov.strip()
    prov = prov.replace("Daerah Istimewa Yogyakarta", "D.I. Yogyakarta")
    return prov


url = "https://catapa.com/blog/daftar-lengkap-kenaikan-ump-2024-di-38-provinsi-indonesia"
resp = requests.get(url)
resp.raise_for_status()
soup = BeautifulSoup(resp.text, "html.parser")
```

```
paras = soup.find_all("p")
text_lines = []
for p in paras:
    text_lines.extend(p.get_text().split("\n"))


pattern = re.compile(r"^([A-Za-z\s\.\-]+)\s*:\s*Rp([\d\.,]+)")

data = []
for line in text_lines:
    m = pattern.match(line)
    if m:
        prov = norm_prov(m.group(1))
        val = rupiah_to_int(m.group(2))
        if prov and val is not None:
            data.append((prov, val))


df_ump = pd.DataFrame(data, columns=["Provinsi", "UMP_2024"])

print("Jumlah provinsi hasil scraping:", len(df_ump))
print(df_ump["Provinsi"].unique())


df_ump.to_csv("ump_2024.csv", index=False)
print("Nama file ump_2024_.csv")

df_ump.head(38)
```

Data UMP 2024 diperoleh melalui *web scraping* dari situs Catapa.com. Kode scraping memanfaatkan library *requests* untuk mengakses halaman web dan *BeautifulSoup* untuk membaca struktur HTML. Nilai UMP kemudian dibersihkan dari format Rupiah menjadi angka murni menggunakan fungsi *rupiah_to_int()*. Hasil scraping disimpan ke dalam Data Frame dan diekspor sebagai ump_2024.csv.

2. Mengekstrak data konsumsi energi dan protein dari PDF BPS.

```
import pdfplumber
import pandas as pd

with pdfplumber.open("1737605184791-81-23.-konsumsi-kalori-dan-protein-penduduk-indonesia-dan-provinsi--maret-2024.pdf") as pdf:
    page = pdf.pages[29]
    table = page.extract_table()

df_kalori = pd.DataFrame(table[1:], columns=table[0])

print(df_kalori)
```

Data konsumsi energi dan protein tersedia dalam format PDF, sehingga diperlukan teknik ekstraksi tabel menggunakan *pdfplumber*. Metode ini mengambil tabel secara langsung dari halaman PDF dan mengubahnya menjadi dataframe.

3. Mengambil data kesehatan dari PDF UNICEF(manual)

```
with pdfplumber.open("Analisis Lanskap Kelebihan Berat Badan dan Obesitas di Indonesia_ Ringkasan Temuan Kunci.pdf") as pdf:
    page = pdf.pages[12]  # Halaman 13 (index mulai dari 0)
    tables = page.extract_tables()
    for table in tables:
        df = pd.DataFrame(table[1:], columns=table[0])
        print(df)
page.debug_tablefinder()
df.head()
```

```
df_kesehatan = pd.read_csv("tabel_gizi_lengkap_part2.csv")
df_kesehatan.head(38)
```

Data kesehatan diperoleh dari laporan PDF UNICEF. File PDF dibaca menggunakan *pdfplumber*, dan tabel-tabel di dalamnya diekstraksi melalui extract_table() . Namun karena struktur tabel pada PDF kompleks dan tidak dapat terbaca dengan baik, oleh karena itu kami menggunakan file CSV hasil input manual yang kami buat *(tabel_gizi_lengkap_part2.csv)* sehingga data yang diperoleh menjadi lebih baik

4. Menggabungkan seluruh dataset menggunakan kolom Provinsi sebagai kunci utama.

```
df_gabungan = df_ump.merge(df_clean_kalori, on="Provinsi", how="left").merge(df_kesehatan, on="Provinsi", how="left")
df_gabungan.head(38)
```

Proses integrasi dilakukan menggunakan fungsi *merge()* Dataset UMP digabungkan dengan dataset konsumsi dan dataset kesehatan. Penggabungan dilakukan dengan metode *left join* untuk membuat seluruh provinsi dari dataset UMP tetap muncul meskipun dari dataset lain tidak tersedia. Supaya dataset lain menyesuaikan dengan provinsi yang berada pada UMP.

E. Teknik Pembersihan Data

1. Membersihkan format angka

```
#mengubah dan membersihkan kolom angka ke numeric
kolom = [
    "Kuintil Pertama", "Kuintil Kedua", "Kuintil Ketiga",
    "Kuintil Keempat", "Kuintil Kelima",
    "Stunting", "Wasting", "Kelebihan berat badan",
    "Kelebihan BB Anak 5-12 tahun", "Kelebihan BB Remaja 13-15 tahun",
    "Kelebihan BB Remaja 16-18 tahun",  "Kelebihan BB Dewasa >18 tahun",
    "Kerawanan pangan"
]

for col in kolom:
    df_gabungan[col] = pd.to_numeric(df_gabungan[col], errors="coerce")

df_gabungan.dtypes
```

Data dari PDF sering mengandung karakter lain seperti titik, koma, spasi, dan huruf yang menyebabkan data tidak dapat dikenali sebagai angka oleh Python sehingga perlu dirubah menjadi numerik agar dapat diproses.

2. Membersihkan data asupan kalori per-kapita menggunakan regex

```
import pandas as pd
import numpy as np
import re

df_clean = df_kalori.copy()

# 1. Hapus baris kosong total
df_clean = df_clean.replace("", np.nan)
df_clean = df_clean.dropna(how="all")

# 2. Set header manual
df_clean.columns = [
    "Provinsi",
    "Kuintil Pertama",
    "Kuintil Kedua",
    "Kuintil Ketiga",
    "Kuintil Keempat",
    "Kuintil Kelima"
]

# 3. Drop baris "(1)" "(2)"
df_clean = df_clean[~df_clean["Provinsi"].str.contains(r"\(", na=False)]

# 4. Bersihkan kolom angka: HAPUS SEMUA HURUF dan simbol selain angka/.,,
angka_cols = ["Kuintil Pertama", "Kuintil Kedua", "Kuintil Ketiga",
              "Kuintil Keempat", "Kuintil Kelima"]

for col in angka_cols:
    df_clean[col] = df_clean[col].astype(str)
    df_clean[col] = df_clean[col].apply(lambda x: re.sub(r"[A-Za-z/\\:\n\s]", "", x))
    # sekarang hanya tersisa "1661,76" atau "1.660,82"

# 5. Convert format Eropa
def europe(x):
    x = x.replace(".", "").replace(",", ".")
    try:
        return float(x)
    except:
        return np.nan

for col in angka_cols:
```

```
    df_clean[col] = df_clean[col].apply(europe)

# 6. Rapikan nama provinsi
df_clean["Provinsi"] = df_clean["Provinsi"].str.replace(r"\d+", "", regex=True)
df_clean["Provinsi"] = df_clean["Provinsi"].str.replace(r"\s+", " ",
regex=True).str.strip()
df_clean = df_clean[df_clean["Provinsi"].str.len() > 2]

df_clean = df_clean.reset_index(drop=True)

df_clean_kalori = df_clean
df_clean_kalori.head(38)
```

Data hasil ekstraksi tabel pada PDF sering kali tidak bersih, banyak object yang tidak diperlukan terbawa, kolom kosong yang terbawa, ataupun format provinsi salah. Maka dari itu perlu adanya proses cleaning sebelum merge ketiga data set.

3. Menyeragamkan nama provinsi

```
#menghapus spasi awal dan akhir kolom provinsi
for df in [df_ump, df_clean_kalori, df_kesehatan]:
    df["Provinsi"] = df["Provinsi"].str.strip()
```

```
    mapping = {
        "DI Yogyakarta": "D.I. Yogyakarta",
        "D I Yogyakarta": "D.I. Yogyakarta",
        "Maluku ": "Maluku",
        "Kep. Bangka Belitung": "Bangka Belitung",
    }
    for df in [df_ump, df_clean_kalori, df_kesehatan]:
        df["Provinsi"] = df["Provinsi"].replace(mapping)
```

Untuk memastikan proses *merge* berjalan sempurna, seluruh nama provinsi diseragamkan menggunakan *mapping dictionary*. Tahap ini sangat penting karena kesalahan kecil dalam penulisan (spasi, titik, huruf kapital) dapat membuat data gagal digabungkan.

4. Menangani missing value

```
#UNTUK PROVINSI PEMEKARAN PAPUA
provinsi_pemekaran_papua = [
    'Papua Pegunungan', 'Papua Barat Daya',
    'Papua Selatan', 'Papua Tengah']

filter_papua = df_gabungan['Provinsi'].isin(provinsi_pemekaran_papua)
filter_non_papua = df_gabungan['Provinsi'].isin(provinsi_pemekaran_papua) == False


# Untuk provinsi pemekaran Papua → samakan dengan Papua induk
nilai_papua_induk = df_gabungan.loc[df_gabungan['Provinsi'] == 'Papua', 'Kerawanan pangan'].values[0]
df_gabungan.loc[filter_papua, 'Kerawanan pangan'] = nilai_papua_induk

# Untuk provinsi lain, isi dengan median nasional
median_nasional = df_gabungan.loc[filter_non_papua, 'Kerawanan pangan'].median()
df_gabungan.loc[df_gabungan['Kerawanan pangan'].isna() & filter_non_papua, 'Kerawanan pangan'] = median_nasional
```

Karena data berasal dari beberapa dokumen yang berbeda maka terdapat beberapa nilai yang kosong (missing), terutama pada indikator kesehatan. Penanganan dilakukan dengan metode imputasi median, yaitu mengisi nilai yang kosong menggunakan nilai tengah dari variabel terkait.

5. Melakukan cek duplikat pada kolom provinsi

```
df_gabungan['Provinsi'].duplicated().sum()
```

Kode ini bertujuan melakukan cek untuk memastikan tidak ada Provinsi yang terhitung dua kali sehingga tidak mempengaruhi hasil akhirnya.

F. Metode Analisis

1. Statistik deskriptif

```
#!.Descriptive statistics
df_clean.describe()
```

Menggunakan fungsi *df_clean.describe()* untuk mengetahui nilai statistika deskiptif yang berupa nilai minimum, maksimum, rata-rata, median, serta standar deviasi dari variabel numerik seperti UMP, konsumsi energi, konsumsi protein dan indikator kesehatan.

2. Distribusi Untuk tiap Kolom

```
plt.figure(figsize=(20,15))
for i, col in enumerate(numeric_cols):
    plt.subplot(4,4,i+1)
    sns.histplot(df_eda[col], kde=True, color='skyblue')
    plt.title(col)
plt.tight_layout()
plt.show()
```

Distribusi untuk tiap kolom atau variabel divisualisasikan menggunakan *sns.histplot(df_clean[col])* untuk membuat histogram yang dapat melihat pola persebaran data, sehingga dapat diketahui apakah data normal, miring atau memiliki outlier.

3. Heatmap atau Korelasi Antar Variabel

```
plt.figure(figsize=(14,10))
sns.heatmap(df_eda[numeric_cols].corr(), annot=True, fmt=".2f",
cmap="coolwarm")
plt.title("Heatmap Korelasi Variabel Numeric")
plt.show()
```

Korelasi dihitung dengan *df_clean.corr()* dan divisualisasikan menggunakan *sns.heatmap()*. Heatmap memudahkan melihat hubungan antar variabel, seperti apakah UMP berkorelasi dengan konsumsi protein atau obesitas.

4. Analisis Provinsi

```
plt.figure(figsize=(12,6))
sns.scatterplot(data=df_eda, x='UMP_2024', y='Stunting', hue='Kerawanan
pangan', size='Kelebihan berat badan', palette='viridis', sizes=(50,300))
plt.title("UMP vs Stunting dengan Kerawanan Pangan & Overweight")
plt.xlabel("UMP 2024")
plt.ylabel("Stunting (%)")
plt.show()
```

sns.scatterplot() digunakan untuk membuat grafik hubungan antar variabel, misalnya UMP dengan stunting atau indikator kesehatan lainnya. Penggunaan parameter seperti size dan hue menambahkan dimensi informasi tambahan dalam satu grafik.

5. Boxplot tiap Kolom

```
for col in numeric_cols:
    plt.figure(figsize=(8,5))
    sns.boxplot(x=df_eda[col], color='lightblue')
    plt.title(f'Boxplot untuk {col}', fontsize=14)
    plt.xlabel(col)
    plt.show()
```

Dengan *sns.boxplot(df_clean[col])*, boxplot menampilkan persebaran data dan mendeteksi nilai ekstrem (outlier). Visualisasi ini membantu apakah ada provinsi dengan nilai sangat tinggi atau rendah pada suatu variabel.

6. Pairplot atau Scatterplot Matrix

```
sns.pairplot(df_eda[numeric_cols], diag_kind='kde')
plt.show()
```

sns.pairplot(df_clean) menampilkan hubungan antar variabel dalam satu tampilan. Pairplot mempermudah identifikasi pola linear maupun non-linear dan memberikan gambaran menyeluruh tentang interaksi variabel.

7. Barplot Top 5 Stunting Tertinggi

```
# Daftar indikator yang mau ditampilkan
metrics = ['Stunting', 'Wasting', 'Kelebihan berat badan',
        'Kelebihan BB Anak 5-12 tahun', 'Kelebihan BB Remaja 13-15 tahun',
            'Kelebihan BB Remaja 16-18 tahun', 'Kelebihan BB Dewasa >18
tahun',
        'Kerawanan pangan']

fig, axes = plt.subplots(4, 2, figsize=(18, 20))
axes = axes.flatten()

for i, col in enumerate(metrics):
   # Ambil top 5 provinsi
                    top5    =    df_eda[['Provinsi',    col]].sort_values(by=col,
ascending=False).head(5)

   sns.barplot(data=top5, x=col, y='Provinsi', ax=axes[i], palette='viridis')
   axes[i].set_title(f'Top 5 Provinsi - {col}')
   axes[i].set_xlabel('')
   axes[i].set_ylabel('')

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

df.nlargest(5, 'kolom') dan sns.barplot() menampilkan provinsi dengan nilai tertinggi pada indikator seperti stunting, wasting, atau obesitas pada berbagai kelompok usia, dan kerawanan pangan. Analisis ini membantu menunjukkan daerah yang paling menonjol atau beresiko.

G. Metode Data Publishing

```
df_ump.to_csv("raw_ump.csv", index=False)
df_clean_kalori.to_csv("raw_kalori.csv", index=False)
df_kesehatan.to_csv("raw_kesehatan.csv", index=False)
df_clean.to_csv("dataset_clean.csv", index=False)
```

Kode ini digunakan untuk menyimpan data mentah yang berupa data ump, data asupan kalori, dan prevalensi kesehatan ke dalam CSV. Selain itu juga menyimpan CSV untuk data yang sudah melalui tahap wrangling. Setelah dijalankan pada folder akan memunculkan dokumen CSV file tersebut.

BAB IV

HASIL DAN PEMBAHASAN

A. Analisis Missing Values

Berdasarkan hasil pemeriksaan kualitas data ditemukan missing values pada beberapa kolom, yaitu kolom kuintil pertama sampai dengan kuintil kelima sebanyak 4 nilai; kolom stunting, wasting, dan kelebihan berat badan sebanyak 14 nilai; kolom Kelebihan BB Anak 5-12 tahun, Kelebihan BB Remaja 13-15 tahun, Kelebihan BB Remaja 16-18 tahun sebanyak 1 nilai; kolom Kelebihan BB Dewasa >18 tahun sebanyak 8 nilai; dan kolom Kerawanan pangan sebanyak 18 nilai. Untuk menangani masalah missing value tersebut dilakukan analisis scatter plot sebelum penanganan missing value dan sesudah penganan. Hasil scatter plot sebelum penanganan menunjukan tidak ada pola linear yang jelas, sehingga metode imputasi berbasis regresi linear kurang tepat. Maka untuk menangani missing value tersebut menggunakan beberapa metode, diantaranya:

1. Menghapus kolom dan baris, dimana jika memiliki missing value lebih dari 50%, maka baris atau kolom tersebut dapat dihapus dari data
2. Mengisi missing values dengan nilai median. Hal ini dilakukan karena median tidak berpengaruh pada outlier. Jika menggunakan rata-rata maka kemungkinan nilai rata-rata akan tertarik oleh nilai outlier
3. Untuk Provinsi pemekaran seperti Papua Pegunungan, Papua Barat Daya, Papua Selatan, Papua Tengah, missing value diganti dengan nilai Provinsi Papua pusat. Hal tersebut dilakukan karena Provinsi Pemekaran Papua memiliki UMP yang sama dengan Provinsi induknya.

	Provinsi	UMP_2024	Kuintil Pertama	Kuintil Kedua	Kuintil Ketiga	Kuintil Keempat	Kuintil Kelima	Stunting	Wasting	Kelebihan berat badan	Kelebihan BB Anak 5-12 tahun	Kelebihan BB Remaja 13-15 tahun	Kelebihan BB Remaja 16-18 tahun	Kelebihan BB Dewasa >18 tahun	Kerawanan pangan
0	Aceh	3460672	1559.07	1863.50	2038.54	2228.84	2574.43	29.3	13.60	3.5	17.2	15.3	12.1	39.9	4.70
1	Bali	2813672	1918.83	2175.64	2266.56	2366.14	2590.63	24.6	10.55	4.7	26.7	20.0	13.2	39.8	4.80
2	Banten	2727812	1752.49	2053.64	2205.32	2428.25	2597.54	23.9	10.20	4.9	19.3	15.6	10.5	36.9	5.76
3	Bengkulu	2507079	1661.21	1943.68	2030.50	2243.65	2420.84	24.6	10.55	4.7	23.6	15.2	7.7	39.4	4.80
4	D.I. Yogyakarta	2125897	1661.76	1935.91	2072.51	2181.61	2412.39	24.6	10.55	4.7	21.6	25.6	19.1	39.0	4.80
5	DKI Jakarta	5067381	1665.07	1987.44	2191.06	2417.09	2528.61	17.6	10.10	6.7	27.3	23.3	19.2	48.0	3.36
6	Gorontalo	3025100	1635.97	1864.76	2014.22	2118.37	2347.49	26.8	12.70	4.4	17.0	14.3	14.8	42.5	5.96
7	Jambi	3037121	1588.60	1837.40	2018.34	2225.61	2543.68	24.6	10.55	4.7	19.2	10.5	5.2	39.4	4.80
8	Jawa Barat	2057495	1682.32	1947.48	2128.39	2293.57	2499.94	21.7	6.30	3.9	18.4	17.0	13.1	39.8	4.90
9	Jawa Tengah	2036947	1579.14	1860.01	2043.44	2207.25	2434.89	24.6	10.55	4.7	21.3	15.2	11.6	36.0	4.80
10	Jawa Timur	2165244	1660.82	1922.42	2061.40	2230.83	2466.43	24.6	10.55	4.7	23.5	19.8	14.9	38.2	4.80
11	Kalimantan Barat	2702616	1476.70	1738.32	1867.23	2118.72	2396.34	24.5	13.30	5.5	20.6	15.5	9.9	30.7	5.02
12	Kalimantan Selatan	3282812	1693.71	1986.93	2177.30	2324.63	2654.93	24.7	12.40	4.8	20.2	14.5	12.9	33.5	3.69
13	Kalimantan Tengah	3261616	1638.19	1910.17	2133.95	2334.62	2652.03	23.5	9.00	7.2	21.4	17.1	12.5	36.5	3.73
14	Kalimantan Timur	3360858	1573.26	1827.89	2017.47	2171.82	2453.21	22.8	10.10	4.7	21.4	18.6	15.7	43.2	3.23
15	Kalimantan Utara	3361653	1465.64	1715.02	1866.90	2042.89	2365.66	17.4	8.70	4.4	24.2	15.8	9.9	40.2	3.65

B. Analisis Outlier

Berdasarkan pemeriksaan kualitas data, ditemukan adanya nilai outlier pada beberapa kolom numerik, terutama pada kolom indikator kesehatan dan UMP provinsi. Outlier ini dapat memengaruhi analisis data dan visualisasi, sehingga perlu dilakukan penanganan. Penanganan outlier dilakukan dengan beberapa metode, yaitu:

1. Menghapus baris dengan outlier banyak
   Baris yang memiliki outlier pada lebih dari 3–4 kolom dihapus dari dataset karena dianggap terlalu ekstrem dan dapat menurunkan kualitas analisis.
2. Mengisi nilai outlier dengan median
   Untuk baris yang hanya memiliki sedikit outlier, nilai outlier digantikan dengan nilai median kolom tersebut.

Median dipilih karena tidak terpengaruh oleh nilai ekstrim, sehingga distribusi data tetap stabil. Hasil boxplot setelah penanganan outlier menunjukkan distribusi data yang lebih representatif dengan sebagian besar kolom sudah tidak memiliki outlier, sehingga dataset siap untuk tahap eksplorasi.

	Provinsi	UMP_2024	Kuintil Pertama	Kuintil Kedua	Kuintil Ketiga	Kuintil Keempat	Kuintil Kelima	Stunting	Wasting	Kelebihan berat badan	Kelebihan BB Anak 5-12 tahun	Kelebihan BB Remaja 13-15 tahun	Kelebihan BB Remaja 16-18 tahun	Kelebihan BB Dewasa >18 tahun	Kerawanan pangan
0	Aceh	3460672	1559.07	1863.50	2038.54	2228.84	2574.43	24.6	13.60	4.7	17.2	15.30	12.1	39.9	4.70
1	Bali	2813672	1918.83	2175.64	2266.56	2366.14	2590.63	24.6	10.55	4.7	26.7	20.00	13.2	39.8	4.80
2	Banten	2727812	1752.49	2053.64	2205.32	2428.25	2597.54	23.9	10.20	4.9	19.3	15.60	10.5	36.9	5.76
3	Bengkulu	2507079	1661.21	1943.68	2030.50	2243.65	2420.84	24.6	10.55	4.7	23.6	15.20	7.7	39.4	4.80
4	D.I. Yogyakarta	2125897	1661.76	1935.91	2072.51	2181.61	2412.39	24.6	10.55	4.7	21.6	15.25	11.1	39.0	4.80
6	Gorontalo	3025100	1635.97	1864.76	2014.22	2118.37	2347.49	26.8	12.70	4.4	17.0	14.30	14.8	42.5	5.96
7	Jambi	3037121	1588.60	1837.40	2018.34	2225.61	2543.68	24.6	10.55	4.7	19.2	10.50	5.2	39.4	4.80
8	Jawa Barat	2057495	1682.32	1947.48	2128.39	2293.57	2499.94	21.7	10.55	4.7	18.4	17.00	13.1	39.8	4.90
9	Jawa Tengah	2036947	1579.14	1860.01	2043.44	2207.25	2434.89	24.6	10.55	4.7	21.3	15.20	11.6	36.0	4.80
10	Jawa Timur	2165244	1660.82	1922.42	2061.40	2230.83	2466.43	24.6	10.55	4.7	23.5	19.80	14.9	38.2	4.80
11	Kalimantan Barat	2702616	1476.70	1738.32	1867.23	2118.72	2396.34	24.5	13.30	4.7	20.6	15.50	9.9	30.7	5.02
12	Kalimantan Selatan	3282812	1693.71	1986.93	2177.30	2324.63	2654.93	24.7	12.40	4.8	20.2	14.50	12.9	33.5	3.69
13	Kalimantan Tengah	3261616	1638.19	1910.17	2133.95	2334.62	2652.03	23.5	9.00	4.7	21.4	17.10	12.5	36.5	3.73
14	Kalimantan Timur	3360858	1573.26	1827.89	2017.47	2171.82	2453.21	22.8	10.10	4.7	21.4	18.60	15.7	43.2	4.80

C. Analisis Statistika Deskriptif

	UMP_2024	Kuintil Pertama	Kuintil Kedua	Kuintil Ketiga	Kuintil Keempat	Kuintil Kelima	Stunting	Wasting	Kelebihan berat badan	Kelebihan BB Anak 5-12 tahun	Kelebihan BB Remaja 13-15 tahun	Kelebihan BB Remaja 16-18 tahun	Kelebihan BB Dewasa >18 tahun	Kerawanan pangan
count	3.400000e+01	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000
mean	3.051832e+06	1579.717353	1850.395882	2013.317059	2200.867059	2503.226471	24.494118	10.841176	4.700000	19.047059	15.242647	11.138235	38.626471	4.817353
std	5.637724e+05	127.255918	132.637101	133.679480	128.979863	109.793343	0.740983	1.600638	0.161433	3.908285	2.120111	3.142284	4.036223	0.646099
min	2.036947e+06	1343.240000	1596.200000	1790.510000	1943.430000	2269.370000	21.700000	7.300000	4.300000	9.900000	10.500000	4.700000	30.700000	3.630000
25%	2.730034e+06	1481.997500	1740.512500	1883.325000	2088.557500	2432.857500	24.600000	10.287500	4.700000	16.225000	14.400000	9.525000	35.850000	4.800000
50%	3.031110e+06	1602.440000	1865.805000	2053.260000	2227.225000	2532.610000	24.600000	10.550000	4.700000	19.000000	15.250000	10.950000	39.400000	4.800000
75%	3.400119e+06	1660.780000	1935.017500	2089.130000	2293.542500	2586.580000	24.600000	11.950000	4.700000	21.400000	15.950000	13.050000	40.650000	4.800000
max	4.024270e+06	1918.830000	2175.640000	2266.560000	2428.250000	2670.340000	26.800000	13.800000	5.200000	28.200000	20.000000	17.200000	47.500000	6.710000

Berdasarkan hasil statistika deskriptif yang dihasilkan ditemukan:

1. UMP_2024
   Berdasarkan hasil analisis statistik, nilai UMP tahun 2024 memiliki 34 observasi dengan nilai rata-rata sebesar 3.051.832, standar deviasi 563.772, nilai minimum 2.036.947, dan maksimum 4.024.270. Distribusi datanya menunjukkan nilai median di 3.031.110 dan kuartil ketiga sebesar 3.400.119, yang mengindikasikan bahwa sebagian besar provinsi memiliki UMP dalam rentang 3 hingga 3.4 juta. Rentang yang cukup lebar antara minimum dan maksimum mengindikasikan adanya variasi tingkat upah antar provinsi di Indonesia.

2. Kuintil Pertama
   Kolom kuintil pertama memiliki 34 nilai dengan rata-rata 1.579,7, standar deviasi 127,26, nilai minimum 1.434,24, dan maksimum 1.918,83. Median berada di 1.602,44, sementara kuartil ketiga adalah 1.660,78. Penyebaran nilai yang tidak terlalu ekstrim menunjukkan bahwa asupan kalori kelompok terbawah antar provinsi cenderung relatif homogen.

3. Kuintil Kedua
   Kuintil kedua memiliki rata-rata 1.850,40 dengan standar deviasi 132,64. Nilai minimum berada pada 1.596,20, sedangkan maksimum 2.175,64. Median tercatat sebesar 1.865,81. Perbedaan antara nilai minimum dan maksimum sedikit lebih besar dibanding kuintil pertama, tetapi masih dalam rentang wajar.

4. Kuintil Ketiga
   Kuintil ketiga memiliki nilai rata-rata sebesar 2.013,32 dan standar deviasi 133,68. Nilai minimum berada pada 1.790,51, sedangkan maksimum mencapai 2.266,56. Median tercatat sebesar 2.035,26 dan kuartil ketiga berada di angka 2.089,13. Variasi yang semakin meningkat dibanding kuintil sebelumnya menunjukkan adanya kesenjangan asupan kalori yang lebih terlihat pada kelompok menengah.

5. Kuintil Keempat
   Kuintil keempat, nilai rata-rata tercatat sebesar 2.200,87 dengan standar deviasi 128,98. Nilai minimum berada pada 1.943,43, sementara nilai maksimum mencapai 2.428,25. Median berada di 2.227,23 dengan kuartil ketiga sebesar 2.293,54. Data ini mengindikasikan bahwa kelompok konsumsi menengah-atas memiliki variasi yang masih stabil meskipun rentangnya mulai lebih lebar.

6. Kuintil Kelima (Asupan Kalori Tertinggi)
   Kuintil kelima menunjukkan nilai rata-rata sebesar 2.503,23 dengan standar deviasi 109,79. Nilai minimum mencapai 2.269,37 dan maksimum 2.670,34. Median adalah 2.532,61 dengan kuartil ketiga sebesar 2.586,88. Rentang nilai yang lebih tinggi mengindikasikan bahwa kelompok konsumsi tertinggi antar provinsi cenderung memiliki tingkat kecukupan kalori yang relatif baik dan stabil.

7. Stunting
   Variabel stunting memiliki 34 observasi dengan rata-rata 24,49 dan standar deviasi 0,74. Nilai minimum adalah 21,70 dan maksimum 26,80. Median tercatat di angka 24,66. Rentang yang sempit menunjukkan bahwa prevalensi stunting antar provinsi relatif konsisten dan masih berada pada tingkat yang mengkhawatirkan.

8. Wasting
   Pada variabel Wasting memiliki nilai rata-rata 10,84 dengan standar deviasi 1,60. Nilai minimum tercatat sebesar 7,30 dan maksimum 13,80. Median berada pada angka 10,55. Data ini menunjukkan adanya perbedaan yang cukup nyata antar provinsi, tetapi masih dalam batas yang umum ditemukan pada wilayah dengan ketimpangan akses pangan.

9. Kelebihan Berat Badan (Umum)
   Variabel kelebihan berat badan memiliki rata-rata 4,70 dengan standar deviasi 0,16. Nilai minimum adalah 4,30 dan maksimum 5,20. Median tercatat sebesar 4,70. Variasi yang kecil menunjukkan bahwa prevalensi overweight pada populasi umum relatif seragam antar provinsi.

10. Kelebihan BB Anak 5–12 Tahun
    Variabel ini memiliki rata-rata 19,04 dengan standar deviasi 3,09. Nilai minimum sebesar 9,90 dan maksimum 28,00. Median berada pada angka 19,00. Rentang nilai yang lebih lebar mengindikasikan adanya perbedaan signifikan kondisi gizi anak antar provinsi.

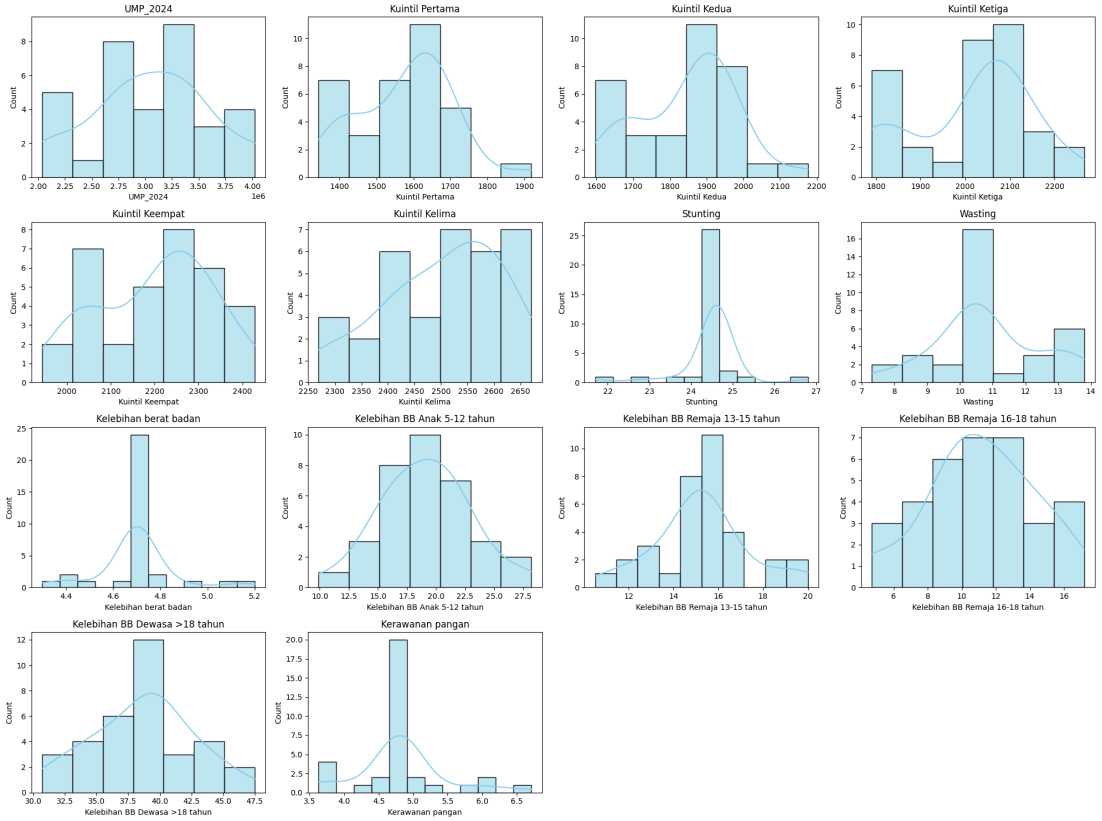
11. Kelebihan BB Remaja 13–15 Tahun
    Rata-rata variabel ini berada pada 15,24 dengan standar deviasi 2,21. Nilai minimum tercatat 10,50 dan maksimum mencapai 20,00. Median adalah 15,25. Variasi antar provinsi cukup besar, menunjukkan adanya ketimpangan kondisi gizi remaja.

12. Kelebihan BB Remaja 16–18 Tahun
    Variabel ini memiliki rata-rata 11,13 dengan standar deviasi 3,14. Nilai minimum adalah 4,70 dan maksimum 17,20. Median berada pada angka 10,95. Rentang variabel yang cukup luas menunjukkan adanya ketimpangan signifikan antar provinsi pada kelompok usia ini.

13. Kelebihan BB Dewasa >18 Tahun
    Rata-rata untuk kelompok ini adalah 38,63 dengan standar deviasi 4,03. Nilai minimum mencapai 30,70 dan maksimum 47,50. Median berada pada 39,40. Angka ini menunjukkan bahwa prevalensi overweight pada orang dewasa cukup tinggi dan bervariasi antar provinsi.

14. Kerawanan Pangan
    Variabel kerawanan pangan memiliki rata-rata sebesar 4,81 dengan standar deviasi 0,64. Nilai minimum tercatat 3,63 dan maksimum 6,71. Median berada pada 4,80. Variasi yang sedang menunjukkan adanya perbedaan tingkat ketahanan pangan antar provinsi, meskipun tidak terlalu ekstrem.

D. Analisis Distribusi Data



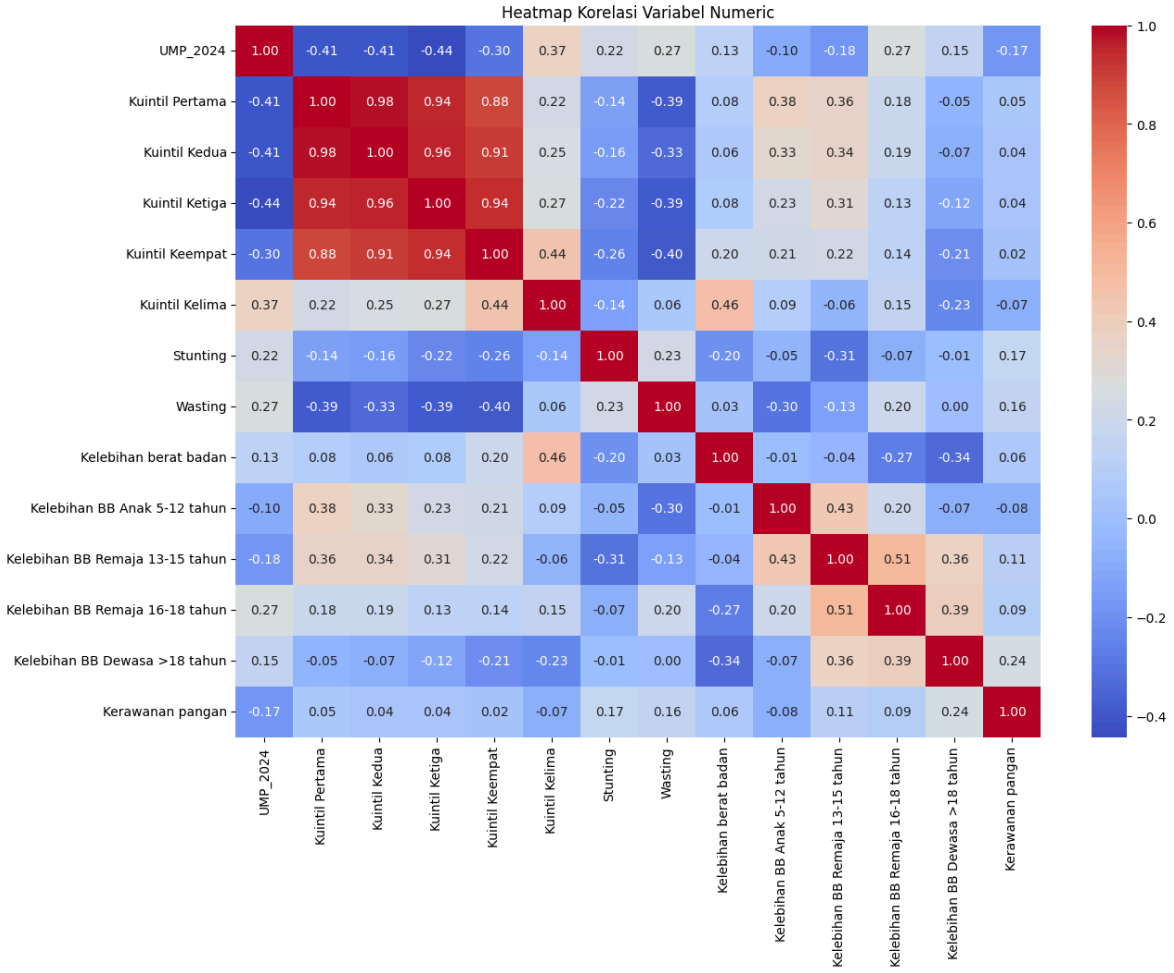
Berdasarkan hasil visualisasi distribusi data, untuk kolom UMP menunjukan data terlihat menyebar cukup merata, namun sedikit condong ke kanan. Mayoritas nilai berada di rentang 2,7–3,4 juta, dengan beberapa provinsi yang memiliki UMP lebih tinggi mendekati 4 juta. Pola ini menunjukkan bahwa sebagian besar provinsi berada pada tingkat upah menengah, sementara hanya sedikit provinsi yang memiliki UMP sangat rendah atau sangat tinggi. Kuantil pertama (asupan kalori kelas terendah) histogram menunjukkan persebaran yang cukup merata, namun cenderung terkonsentrasi pada nilai 1.450–1.750 kalori. Artinya asupan kalori kelompok terendah relatif homogen, dengan sedikit provinsi yang memiliki asupan kalori yang sangat rendah. Untuk kuantil kedua distribusi histogram lebih menyebar dibanding kuintil pertama. Sebagian besar nilai berada di 1.700–1.950 kalori, dengan pola yang mulai membentuk kurva normal. Ini menandakan variasi asupan kalori kelompok ini sedikit lebih besar. Kuintil ketiga mengalami pergeseran distribusi ke arah yang lebih tinggi, mayoritas berada pada rentang 1.850–2.100 kalori. Pola distribusi semakin mirip normal, menandakan kelompok ini lebih stabil dan tidak banyak nilai ekstrim.

Pada histogram kuintil keempat terlihat pola lebih simetris, dengan konsentrasi di 2.050–2.250 kalori. Sedangkan variasi antar provinsi relatif kecil, menunjukkan kelompok ini relatif konsisten dalam asupan kalori. Pada kuantil kelima histogram menunjukkan persebaran nilai yang berkumpul di kisaran 2.400–2.650 kalori. Hampir seluruh daerah memiliki nilai tinggi yang konsisten, menunjukkan bahwa kelompok paling tinggi memiliki tingkat asupan kalori paling baik.

Visualisasi histogram stunting sangat mengerucut pada nilai di 24–25%, menunjukkan penyebaran data sangat sempit. Sehingga, tingkat stunting antarprovinsi cenderung tidak jauh berbeda, meskipun tetap berada pada angka yang cukup tinggi. Sedangkan untuk distribusi wasting terlihat cukup menyebar, namun tetap terkonsentrasi pada kisaran 10–12%. Namun, pada beberapa provinsi memiliki nilai lebih rendah (sekitar 7–9%), menunjukkan adanya variasi nilai wasting. Histogram kelebihan berat badan menunjukan mayoritas nilai berada pada 4,6%-4,8% menunjukan bahwa distribusi sangat sempit.

Berdasarkan visualisasi histogram Kelebihan Berat Badan Menurut Kelompok Umur, untuk histogram anak 5–12 Tahun menunjukkan pola mirip normal, dengan rentang nilai di 16–22%. Terdapat variasi yang cukup jelas antar provinsi, menandakan risiko overweight pada anak usia sekolah berbeda-beda. Pada remaja 13–15 Tahun distribusi berada pada rentang 14–17%, dengan beberapa nilai ekstrem di bawah 12 atau di atas 18. Variasi ini menunjukkan perbedaan resiko overweight pada remaja antar provinsi. Pada remaja 16–18 Tahun sebaran data stabil, dimana mayoritas berada di 9–13%. Kurva menunjukkan pola yang hampir normal, menandakan perbedaan risiko overweight lebih kecil pada kelompok ini. Sedangkan untuk dewasa (>18 Tahun) memiliki variasi paling lebar. Nilai tersebar di rentang 30–47%, menunjukkan bahwa overweight pada orang dewasa sangat berbeda antar provinsi. Beberapa daerah memiliki prevalensi overweight yang jauh lebih tinggi. Untuk distribusi kerawanan pangan condong ke kanan, dengan banyak provinsi berada pada tingkat kerawanan 4–5. Namun ada beberapa provinsi dengan nilai lebih tinggi hingga 6,7, sehingga menunjukkan adanya provinsi dengan kerentanan pangan lebih serius.

E. Analisis Korelasi antar Variabel



Berdasarkan hasil visualisasi korelasi menggunakan heatmap, terlihat bahwa korelasi antar-kolom kuintil (Kuintil 1 hingga Kuintil 5) memiliki nilai yang sangat tinggi, bahkan banyak yang mendekati 0.90 hingga 1.00. Korelasi yang sangat kuat ini merupakan hal yang wajar karena kuintil pada dasarnya mengukur fenomena yang sama, yaitu rata-rata asupan kalori berdasarkan kelompok ekonomi dalam satu provinsi. Apabila sebuah provinsi memiliki tingkat asupan kalori tinggi pada kelompok ekonomi termiskin (kuintil pertama), maka umumnya provinsi tersebut juga menunjukkan asupan kalori yang tinggi pada kelompok ekonomi di atasnya. Dengan demikian, kelima kuintil bergerak mengikuti pola yang sama, sehingga menghasilkan korelasi yang sangat besar dan menunjukkan bahwa variabel-variabel ini tidak independen satu sama lain.

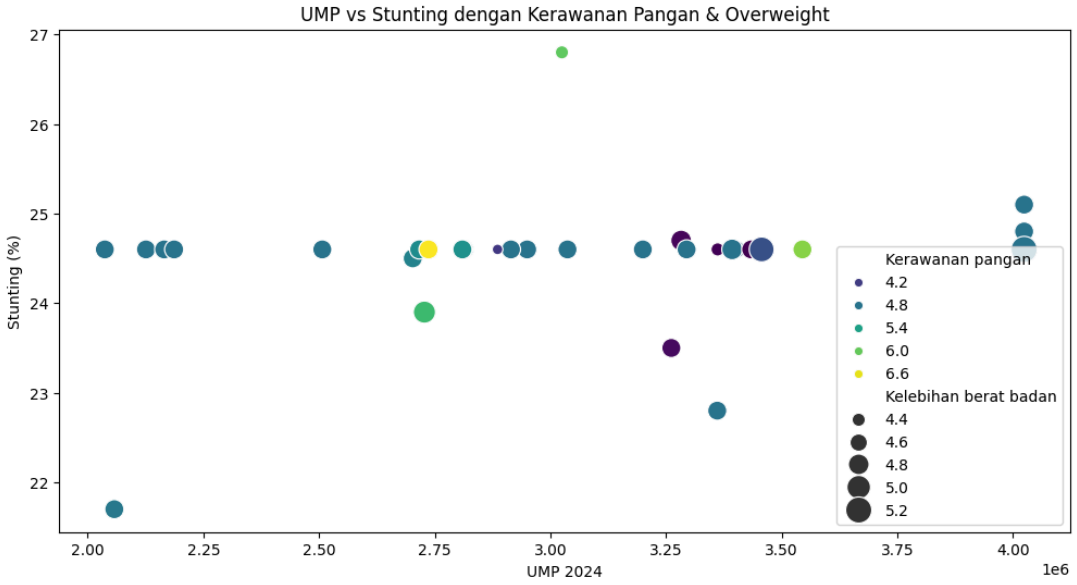
Selain itu UMP dengan indikator kesehatan yang berupa stunting, wasting, atau kelebihan berat badan memiliki korelasi namun tidak kuat satu sama lain, karena heatmap menunjukan pada angka mendekati nol. Hal ini menunjukan bahwa provinsi dengan UMP tinggi tidak selalu memiliki angka stunting yang rendah, begitupun dengan provinsi yang memiliki UMP rendah tidak selalu memiliki angka stunting/wasting tinggi. Asupan kalori dalam kuintil juga menunjukan korelasi yang lemah stunting dan wasting . Hal ini menunjukan bahwa tinggi rendahnya asupan kalori tidak berkaitan secara kuat dan langsung terhadap stunting dan wasting. Namun

korelasi antara kuintil kelima dengan berat badan menunjukan korelasi positif yang cukup kuat dengan angka 0,46. Hal ini menunjukan bahwa asupan konsumsi kelompok ekonomi atas mempengaruhi kelebihan berat badan semakin tinggi.

Korelasi antara UMP dan kerawanan pangan juga menunjukan hubungan yang tidak kuat. hal ini menunjukan bahwa tingkat upah minimum provinsi tidak secara langsung mempengaruhi tingkat ketahan pangan suatu provinsi.

Korelasi antara kelebihan berat badan yang dibagi menjadi beberapa kelompok usia menunjukkan korelasi positif yang cukup kuat dalam rentang 0,24–0,51 terhadap satu sama lain. Hal ini menunjukan bahwa kelebihan berat badan saat usia anak-anak (5-12 tahun) dapat mempengaruhi terhadap kelebihan berat badan usia 13–15 tahun, dan berlanjut hingga kelebihan berat badan dewasa (umur >18 tahun). Jadi, jika saat usia anak-anak memiliki kelebihan berat badan, maka memiliki kelebihan berat badan pada usia selanjutnya berpotensi besar.

F. Analisis Provinsi



Hasil visualisasi menunjukan hubungan antara UMP dan angka stunting tidak bersifat linear. Banyak provinsi dengan UMP tinggi berkisar antara 3 hingga 4 juta rupiah tetap memiliki tingkat stunting di kisaran 24 hingga 25 persen. Sebaliknya, provinsi dengan UMP rendah sekitar 2 juta rupiah pun menunjukkan tingkat stunting yang hampir sama, terlihat satu provinsi dengan angka stunting mencapai sekitar 27 persen meskipun memiliki UMP sekitar 3 juta rupiah hal ini menunjukkan bahwa peningkatan UMP tidak secara otomatis berkorelasi dengan penurunan stunting.

Warna titik pada grafik menunjukkan tingkat kerawanan pangan, di mana provinsi dengan warna lebih terang memiliki kerawanan pangan lebih tinggi. Provinsi-provinsi ini umumnya memiliki angka stunting lebih besar dibanding provinsi lain dengan UMP serupa, sehingga ketahanan pangan terlihat lebih mempengaruhi tingginya stunting dibanding faktor ekonomi semata. Selain itu,

ukuran titik menggambarkan prevalensi overweight. Provinsi dengan UMP tinggi cenderung memiliki ukuran titik lebih besar, menandakan tingkat overweight yang lebih tinggi, sedangkan provinsi ber-UMP rendah memiliki overweight lebih kecil. Pola ini menunjukkan fenomena *double burden of malnutrition*, yaitu adanya beban ganda berupa stunting pada kelompok ekonomi rendah dan overweight pada kelompok ekonomi tinggi.

Secara keseluruhan, sebaran titik pada grafik menunjukkan bahwa hampir seluruh provinsi memiliki tingkat stunting yang relatif berdekatan, yaitu sekitar 24 hingga 25 persen, sehingga masalah stunting dapat dikatakan bersifat nasional dan bukan hanya terjadi pada wilayah tertentu. Hal ini menunjukan bahwa peningkatan pendapatan saja tidak cukup untuk memperbaiki status gizi tanpa adanya perbaikan ketahanan pangan, pemerataan akses pangan bergizi, edukasi gizi dan layanan kesehatan masyarakat yang memadai.

G. Analisis Barplot Top 5 Provinsi

Berdasarkan hasil barplot untuk data kolom “Stunting” menunjukan bahwa 5 provinsi dengan stunting tertinggi adalah Gorontalo dengan persentase lebih dari 25%, kemudian Papua Pegunungan, Papua, Kalimantan Selatan, dan terakhir adalah bali yang nilainya mendekati 25%.

Barplot untuk Wasting menunjukan 5 provinsi dengan nilai kurang gizi akut tertinggi. Provinsi tertinggi pertama yaitu Papua Selatan, kemudian Aceh, diikuti Papua Barat daya, Kalimantan Barat, dan terakhir Papua.

Provinsi dengan kelebihan berat badan tertinggi yaitu Papua Selatan, kemudian Sumatera Selatan dengan persentase lebih dari 5%. Untuk persentase kurang dari 5% menunjukan provinsi Banten, Kalimantan Selatan, dan Papua Barat.

Pada barplot kelebihan berat badan pada anak umur 5-12 tahun didapatkan, provinsi dengan peringkat tertinggi yaitu Papua Pegunungan kemudian bali dengan persentase lebih dari 25%. Diikuti oleh Kalimantan Barat, Bengkulu, dan Jawa Timur dengan persentase dibawah 25%.

Provinsi dengan kelebihan berat badan pada remaja usia 13–15 tahun peringkat pertama adalah Bali dengan persentase 20%. Kemudian Jawa Timur dengan persentase mendekati 20%. Sulawesi Utara, Kalimantan Timur, dan Kalimantan Tengah dengan persentase antara 16-18%.

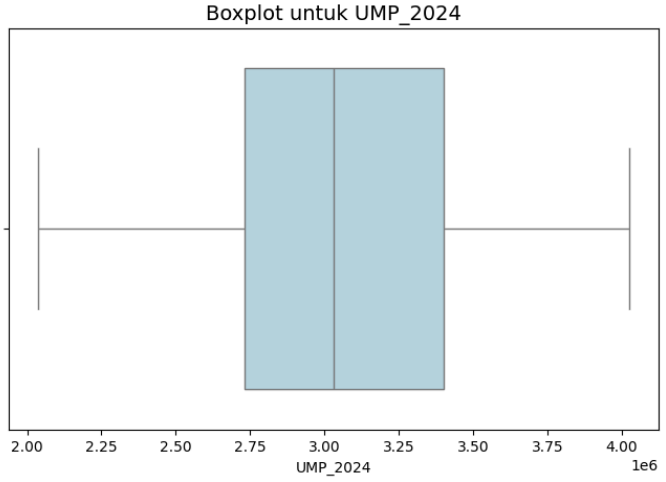
Peringkat pertama kelebihan berat badan pada remaja usia 16-18 tahun adalah Papua Barat Daya dengan persentase kurang lebih 16%. Kemudian diikuti oleh Kepulauan Riau, Kalimantan Timur, Papua, dan Jawa TImur.

Kelebihan berat badan pada orang dewasa yang berumur lebih dari 18 tahun, peringkat pertama dimiliki oleh Sulawesi Selatan, lalu peringkat dua yaitu Papua, peringkat tiga yaitu Kepulauan Riau, peringkat empat yaitu Papua Barat Daya, peringkat lima yaitu Papua Barat. Persentase kelima provinsi tersebut yaitu lebih dari 40%.

Kerawanan pangan berada persentase yang cukup kecil yaitu dibawah 7%. untuk peringkat tertinggi dimiliki oleh Sulawesi tengah, peringkat dua Sulawesi Utara, peringkat tiga Gorontalo, peringkat empat Banten, peringkat kelima yaitu Sumatera Utara.

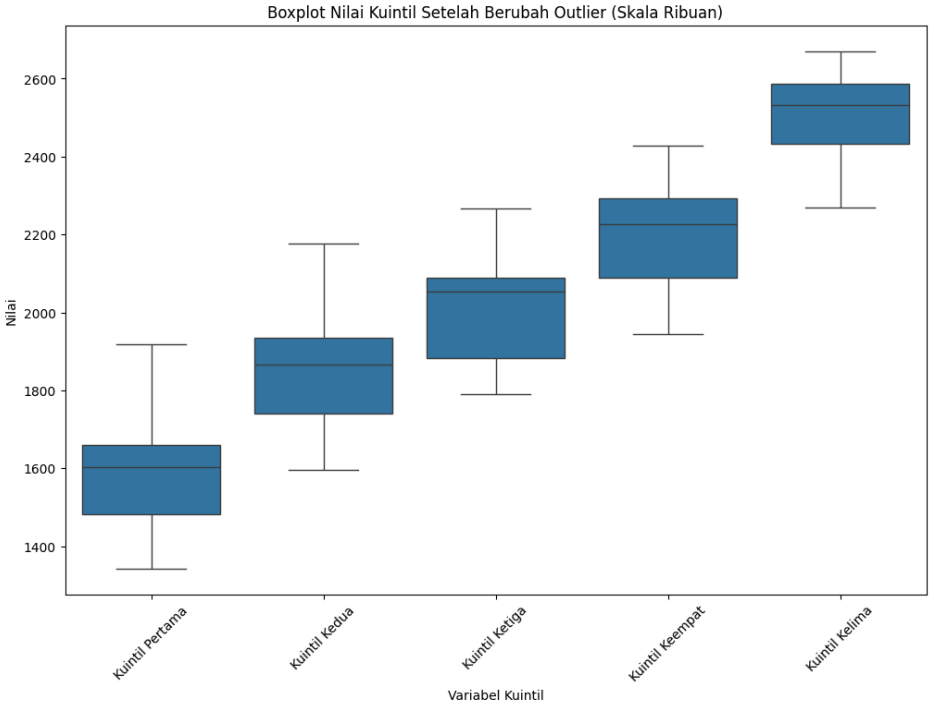
H. Analisis Boxplot tiap Kolom

1. Boxplot UMP 2024



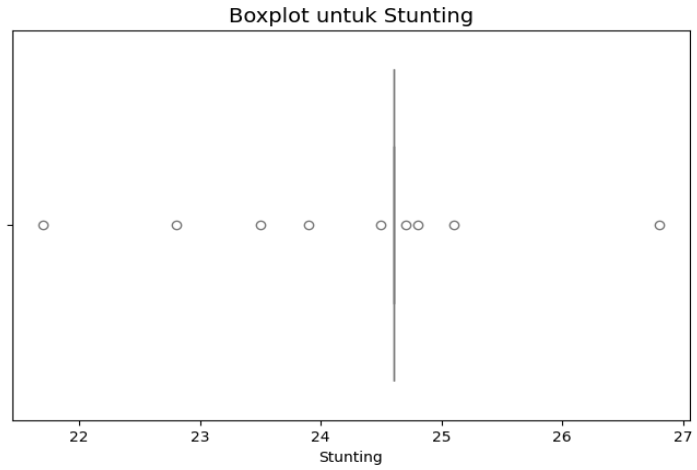
Pada boxplot UMP ini menunjukan bahwa data tersebar antara 2,0 hingga 4,0 (satuaan juta). hal itu menunjukan sebaran data cukup besar, dan tidak terdapat Outlier untuk data UMP.

2. Boxplot Kuintil pertama–Kuintil kelima



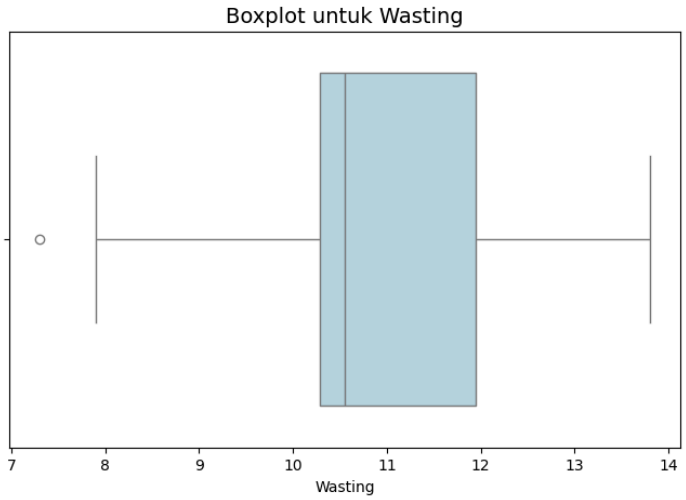
Pada boxplot asupan kalori per-kuantil menunjukan bahwa pola semakin tinggi kuantil maka semakin tinggi nilai minimum dan maksimum asupan kalori. Kuintil dengan persebaran data paling besar berada pada kuintil ketiga. Dari kelima boxplot tersebut nilai asupan kalori per-kuintil tidak memiliki outlier

3. Boxplot Stunting



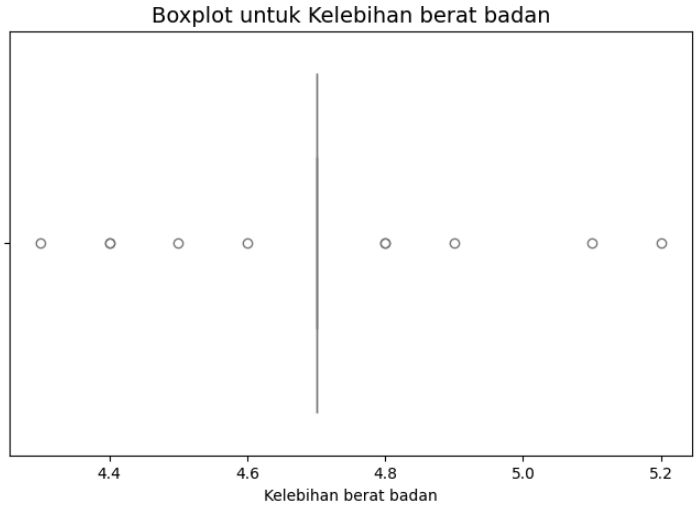
Pada boxplot stunting terlihat bahwa terdapat beberapa nilai outlier yang tidak signifikan, dan data tersebar pada rentang sekitar 22% hingga 27%. Sebarannya sempit menunjukkan bahwa angka stunting antar provinsi relatif seragam.

4. Boxplot wasting



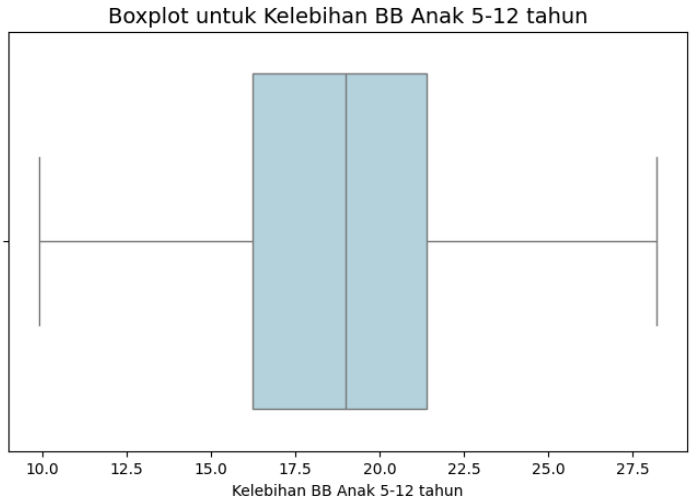
Boxplot wasting menunjukkan variasi data yang lebih lebar, berada pada kisaran 7% hingga 14%, serta terdapat satu titik outlier pada nilai rendah. Hal ini menunjukkan perbedaan wasting yang cukup signifikan antarprovinsi.

5. Boxplot Kelebihan berat badan



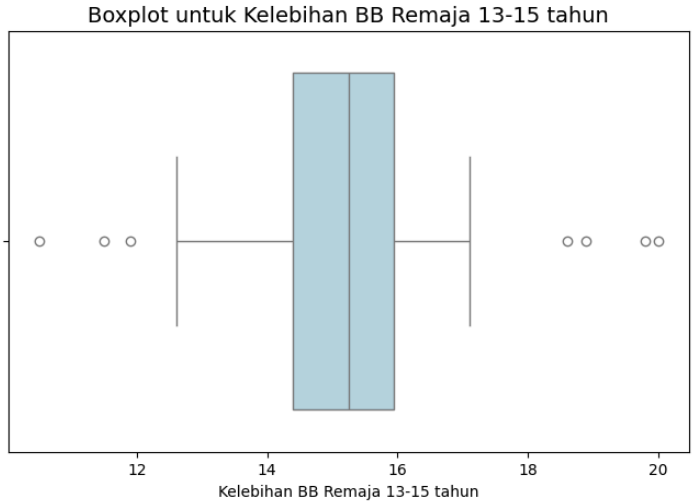
Pada boxplot overweight, tidak terlihat outlier dan data tersebar sempit antara 4,3% hingga 5,2%, menandakan kestabilan angka overweight antarprovinsi.

6. Boxplot Kelebihan BB Anak 5-12 tahun



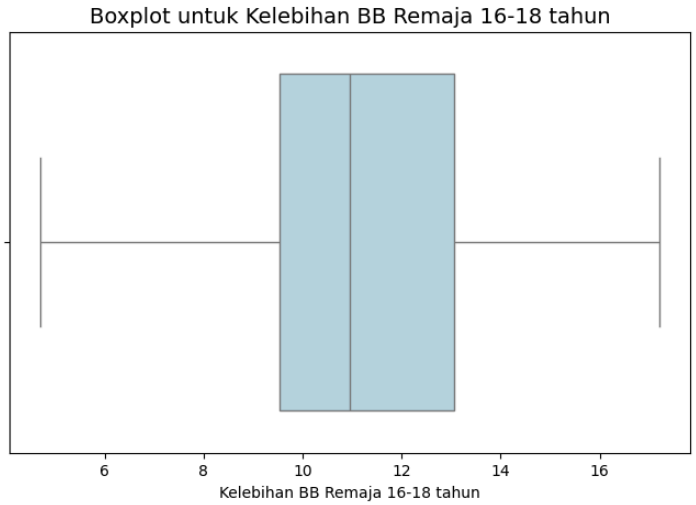
Boxplot ini memperlihatkan rentang data yang cukup luas antara 10% hingga 28%, tanpa outlier ekstrim. Variasi yang besar menunjukkan ketimpangan overweight pada anak sekolah di berbagai provinsi.

7. Boxplot Kelebihan BB Remaja 13-15 tahun



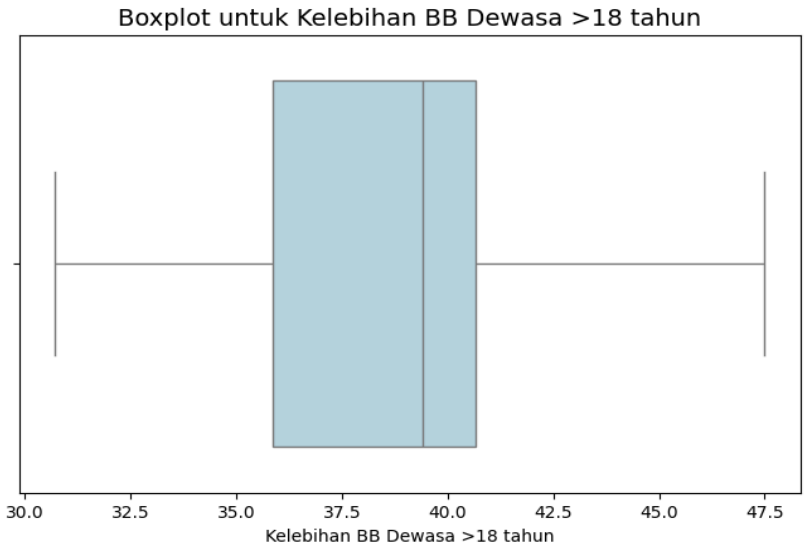
Pada remaja 13–15 tahun, data overweight tersebar antara 11% sampai 20%, dan beberapa provinsi membentuk outlier pada bagian atas. Ini menunjukkan adanya provinsi dengan overweight remaja jauh lebih tinggi dari rata-rata.

8. Boxplot Kelebihan BB Remaja 16-18 tahun



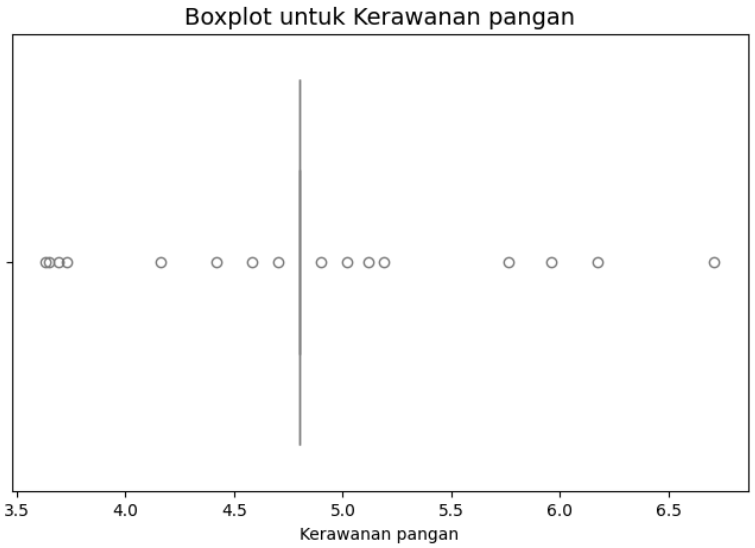
Boxplot ini menunjukkan data tersebar antara 5% hingga 17% tanpa outlier yang ekstrem. Variasinya cukup lebar, menunjukkan perbedaan signifikan status gizi remaja antar provinsi.

9. Boxplot Kelebihan BB Dewasa >18 tahun



Boxplot menunjukkan bahwa overweight pada dewasa memiliki sebaran cukup besar antara 30% hingga 48% tanpa outlier ekstrim. Angka overweight dewasa relatif tinggi di hampir semua provinsi.

10. Boxplot Kerawanan pangan



Pada boxplot kerawanan pangan, data tersebar pada rentang 3,6% hingga 6,7% dan terdapat beberapa titik yang menjauh dari median. Ini menunjukkan nilai sekitar 3.6%–3.8% (lebih rendah) dan 6.0%–6.7% (lebih tinggi) terbaca sebagai outlier

BAB V

KESIMPULAN

A. Kesimpulan

Berdasarkan serangkaian proses data wrangling yang telah dilakukan, dapat disimpulkan bahwa data yang didapatkan dari suatu sumber tidak bisa langsung dianalisis, karena data belum tentu memiliki kualitas yang baik. Data awal mengalami beberapa masalah awal yaitu berupa missing value, format file tidak rata, sumber data berbeda, terdapat objek tidak diperlukan, ataupun outlier. Hal tersebut harus ditangani sesuai dengan data yang yang diambil. Missing value missing values terdapat pada kolom asupan kalori (Kuintil 1–5), prevalensi gizi, dan kerawanan pangan.

Missing values ditangani menggunakan beberapa pendekatan:

1. menghapus baris atau kolom dengan persentase missing value yang terlalu besar
2. mengisi nilai hilang menggunakan median karena lebih stabil terhadap outlier, dan
3. mengisi dengan nilai Papua induk untuk Provinsi pemekaran Papua.

Outlier ditangani menggunakan kombinasi penghapusan baris (jika memiliki outlier lebih dari tiga) dan penggantian nilai ekstrim menggunakan median, agar distribusi data menjadi lebih konsisten. Setelah proses pembersihan, dataset menjadi lebih seragam dan tidak memiliki missing values.

Hasil eksplorasi data (EDA) menunjukkan bahwa Kuintil 1–5 memiliki korelasi sangat tinggi (mendekati 1), menandakan bahwa pola asupan kalori antar kelompok ekonomi dalam satu provinsi bergerak secara konsisten. Sementara itu, UMP, asupan kalori, indikator gizi, dan kerawanan pangan memiliki korelasi yang rendah, sehingga tidak menunjukkan adanya hubungan linear yang kuat. Hal ini mengindikasikan bahwa faktor ekonomi tidak serta-merta memiliki hubungan langsung dengan masalah gizi di provinsi-provinsi Indonesia.

Secara keseluruhan, proses data wrangling berhasil menghasilkan dataset yang bersih, terstruktur, dan siap digunakan untuk analisis lebih dalam.

B. Kendala Data dan Teknis

Pada pengerjaan proyek data wrangling ini tidak terlepas dari berbagai kendala. Kendala yang kami alami berupa:

1. Data yang bersumber dari PDF laporan prevalensi kesehatan UNICEF dan Kemenkes memiliki bentuk yang sangat kompleks, sehingga sangat sulit dilakukan ekstraksi menggunakan metode yang ada. Untuk mengatasi masalah itu kami membuat manual data dengan memasukan nilai pada PDF ke dalam CSV
2. Sulit menemukan data yang relevan dengan kasus.
3. Keterbatasan data yang tersedia terutama untuk data provinsi terbaru

4. Dengan keterbatasan data membuat banyak kolom yang memiliki nilai NaN yang perlu diatasi.
5. Perbedaan Skala antar variabel. Variabel UMP dalam skala jutaan rupiah, sedangkan kuintil dalam skala ribuan kalori. Berbanding terbalik dengan data prevalensi kesehatan yang berupa persentase.
6. Menentukan Analisis yang cocok untuk menyajikan data dengan informatif.

C. Rencana analisis lanjutan

Untuk mengembangkan analisis, maka perlu beberapa tahapan lanjutan untuk mendapatkan pemahaman lebih terkait hubungan antara UMP, asupan kalori, prevalensi kesehatan, dan kerawanan pangan. Tahapan lanjutan dapat berupa:

1. Menambahkan dataset tahun-tahun sebelumnya untuk analisis tren ataupun menambahkan variabel yang dapat menunjukan korelasi kuat antara variabel UMP, asupan kalori, dan prevalensi kesehatan
2. Membuat *geospatial mapping* (peta Indonesia) untuk memvisualisasikan distribusi stunting, wasting, dan kerawanan pangan per provinsi.
3. Analisis hubungan kausal dan prediktif menggunakan decision tree atau random forest.
4. Melakukan *clustering* seperti menggunakan algoritma KNN untuk mengelompokkan provinsi berdasarkan kombinasi indikator kalori, kesehatan, UMP, dan kerawanan pangan.

D. Kontribusi Anggota

- Salsabila Alika Seftizianka (117): Sumber data, Kode (Teknik pengambilan data dan integrasi, Cleaning, Eksplorasi), Laporan, PPT
- Naufal Muzaki (061): Sumber Data, Kode (Teknik pengambilan data, cleaning), Laporan, PPT

DAFTAR PUSTAKA

Badan Pusat Statistik. (2024). *Konsumsi Kalori dan Protein Penduduk Indonesia dan Provinsi, Maret 2024*. Jakarta: BPS. (PDF: 1737605184791-81-23.-konsumsi-kalori-dan-protein-penduduk-indonesia-dan-provinsi--maret-2024.pdf)

Catapa. (2024). *Daftar Lengkap Kenaikan UMP 2024 di 38 Provinsi Indonesia*. Diakses dari https://catapa.com/blog/daftar-lengkap-kenaikan-ump-2024-di-38-provinsi-indonesia

UNICEF Indonesia. (2023). *Analisis Lanskap Kelebihan Berat Badan dan Obesitas di Indonesia: Ringkasan Temuan Kunci*. Jakarta: UNICEF Indonesia. (PDF: Analisis Lanskap Kelebihan Berat Badan dan Obesitas di Indonesia_ Ringkasan Temuan Kunci.pdf)

Durrotun, N. D. K. (2024). *Pengaruh Pertumbuhan Ekonomi, Pendidikan dan Upah Minimum Provinsi Terhadap Kesejahteraan Masyarakat di Indonesia*. Jurnal Ilmu Ekonomi (JIE), 8(02). https://doi.org/10.22219/jie.v8i02.32853

Hasan, F. E. (2015). *Hubungan Tingkat Pendapatan Keluarga, Konsumsi Energi dan Protein dengan Status Gizi Karyawan Tambang Nikel*. MyJurnal Poltekkes Kendari. https://myjurnal.poltekkes-kdi.ac.id/index.php/hijp/article/view/530

Ningsih, C., & Masrikhiyah, R. (2021). *Hubungan Pendapatan, Tingkat Pendidikan dan Tingkat Kecukupan Energi terhadap Status Gizi Ibu Hamil*. JIGK – Jurnal Ilmu Gizi dan Kesehatan. https://jurnal.umus.ac.id/index.php/JIGK/article/view/566

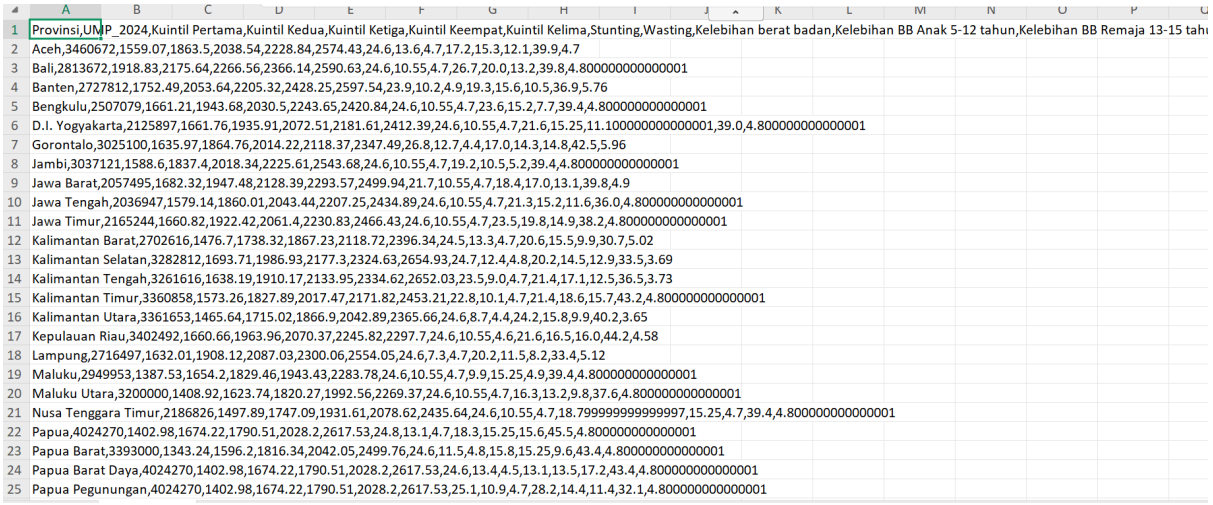
Naibaho, E., & Aritonang, E. (2021). *Pendapatan dan Pengetahuan Gizi Ibu dengan Ketahanan Pangan Keluarga*. *Trophico Journal*, Universitas Sumatera Utara. https://talenta.usu.ac.id/trophico/article/view/8654

LAMPIRAN

Lampiran 1: Publikasi Data Wrangling

- Nama:
  Data-Wrangling-Sumber-Data-UMP-Asupan-Kalori-dan-Prevalensi-Kesehatan.
- Link github:
  [Data-Wrangling-Sumber-Data-UMP-Asupan-Kalori-dan-Prevalensi-Kesehatan.]

Lampiran 2: Dataset bersih



	A
1	Provinsi,UMP_2024,Kuintil Pertama,Kuintil Kedua,Kuintil Ketiga,Kuintil Keempat,Kuintil Kelima,Stunting,Wasting,Kelebihan berat badan,Kelebihan BB Anak 5-12 tahun,Kelebihan BB Remaja 13-15 tah
2	Aceh,3460672,1559.07,1863.5,2038.54,2228.84,2574.43,24.6,13.6,4.7,17.2,15.3,12.1,39.9,4.7
3	Bali,2813672,1918.83,2175.64,2266.56,2366.14,2590.63,24.6,10.55,4.7,26.7,20.0,13.2,39.8,4.800000000000001
4	Banten,2727812,1752.49,2053.64,2205.32,2428.25,2597.54,23.9,10.2,4.9,19.3,15.6,10.5,36.9,5.76
5	Bengkulu,2507079,1661.21,1943.68,2030.5,2243.65,2420.84,24.6,10.55,4.7,23.6,15.2,7.7,39.4,4.800000000000001
6	D.I. Yogyakarta,2125897,1661.76,1935.91,2072.51,2181.61,2412.39,24.6,10.55,4.7,21.6,15.25,11.100000000000001,39.0,4.800000000000001
7	Gorontalo,3025100,1635.97,1864.76,2014.22,2118.37,2347.49,26.8,12.7,4.4,17.0,14.3,14.8,42.5,5.96
8	Jambi,3037121,1588.6,1837.4,2018.34,2225.61,2543.68,24.6,10.55,4.7,19.2,10.5,5.2,39.4,4.800000000000001
9	Jawa Barat,2057495,1682.32,1947.48,2128.39,2293.57,2499.94,21.7,10.55,4.7,18.4,17.0,13.1,39.8,4.9
10	Jawa Tengah,2036947,1579.14,1860.01,2043.44,2207.25,2434.89,24.6,10.55,4.7,21.3,15.2,11.6,36.0,4.800000000000001
11	Jawa Timur,2165244,1660.82,1922.42,2061.4,2230.83,2466.43,24.6,10.55,4.7,23.5,19.8,14.9,38.2,4.800000000000001
12	Kalimantan Barat,2702616,1476.7,1738.32,1867.23,2118.72,2396.34,24.5,13.3,4.7,20.6,15.5,9.9,30.7,5.02
13	Kalimantan Selatan,3282812,1693.71,1986.93,2177.3,2324.63,2654.93,24.7,12.4,4.8,20.2,14.5,12.9,33.5,3.69
14	Kalimantan Tengah,3261616,1638.19,1910.17,2133.95,2334.62,2652.03,23.5,9.0,4.7,21.4,17.1,12.5,36.5,3.73
15	Kalimantan Timur,3360858,1573.26,1827.89,2017.47,2171.82,2453.21,22.8,10.1,4.7,21.4,18.6,15.7,43.2,4.800000000000001
16	Kalimantan Utara,3361653,1465.64,1715.02,1866.9,2042.89,2365.66,24.6,8.7,4.4,24.2,15.8,9.9,40.2,3.65
17	Kepulauan Riau,3402492,1660.66,1963.96,2070.37,2245.82,2297.7,24.6,10.55,4.6,21.6,16.5,16.0,44.2,4.58
18	Lampung,2716497,1632.01,1908.12,2087.03,2300.06,2554.05,24.6,7.3,4.7,20.2,11.5,8.2,33.4,5.12
19	Maluku,2949953,1387.53,1654.2,1829.46,1943.43,2283.78,24.6,10.55,4.7,9.9,15.25,4.9,39.4,4.800000000000001
20	Maluku Utara,3200000,1408.92,1623.74,1820.27,1992.56,2269.37,24.6,10.55,4.7,16.3,13.2,9.8,37.6,4.800000000000001
21	Nusa Tenggara Timur,2186826,1497.89,1747.09,1931.61,2078.62,2435.64,24.6,10.55,4.7,18.799999999999997,15.25,4.7,39.4,4.800000000000001
22	Papua,4024270,1402.98,1674.22,1790.51,2028.2,2617.53,24.8,13.1,4.7,18.3,15.25,15.6,45.5,4.800000000000001
23	Papua Barat,3393000,1343.24,1596.2,1816.34,2042.05,2499.76,24.6,11.5,4.8,15.8,15.25,9.6,43.4,4.800000000000001
24	Papua Barat Daya,4024270,1402.98,1674.22,1790.51,2028.2,2617.53,24.6,13.4,4.5,13.1,13.5,17.2,43.4,4.800000000000001
25	Papua Pegunungan,4024270,1402.98,1674.22,1790.51,2028.2,2617.53,25.1,10.9,4.7,28.2,14.4,11.4,32.1,4.800000000000001