

Tugas Besar Analisis Data Statistik

Ukuran pemusatan dan penyebaran data; Eksplorasi data (dot plot, stem leaf, histogram, boxplot, skewness, deteksi outlier)



Disusun Oleh:

Rafli Al Mansyah Tambunan (124450007)

Fitra Pratama Mega (124450070)

Muhammad Rizaldi (124450093)

Yulia Kristine Malau (124450119)

**KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN
TEKNOLOGI INSTITUT TEKNOLOGI SUMATERA
2025**

Ringkasan Eksekutif

Dalam laporan ini, fokus analisis beralih ke data antropometri mahasiswa Institut Teknologi Sumatera, yaitu Tinggi Badan, Berat Badan, dan Indeks Massa Tubuh (IMT). Analisis dilakukan untuk menghitung dan mengeksplorasi ukuran pemusatan dan penyebaran data, serta memvisualisasikan distribusi dan status gizi mahasiswa. Metode yang digunakan adalah Statistik Deskriptif (Mean, Median, Standar Deviasi, Skewness) dan Eksplorasi Data Visual (Histogram, Box Plot, Bar Chart).

Latar Belakang

Tujuan utama dari tugas besar mata kuliah Analisis Data Statistik ini adalah menunjukkan kompetensi mahasiswa dalam mengimplementasikan metode statistik pada kasus dunia nyata, khususnya melalui penggunaan data Tinggi Badan dan Berat Badan untuk menghitung Indeks Massa Tubuh (IMT) dan mengkategorikan status gizi. Secara spesifik, penelitian ini bertujuan untuk mendemonstrasikan kemampuan mahasiswa dalam pemahaman metode statistik, terutama ukuran pemusatan dan penyebaran, membersihkan data dan menghitung variabel turunan (IMT), serta menunjukkan kemahiran dalam penggunaan bahasa pemrograman R untuk analisis data.

Metodologi

Metode analisis utama yang digunakan dalam penelitian ini adalah Statistik Deskriptif dan Eksplorasi Data Visual. Metode ini dipilih untuk memperoleh pemahaman awal yang komprehensif mengenai distribusi dan karakteristik data Tinggi Badan, Berat Badan, dan Indeks Massa Tubuh (IMT) mahasiswa. Statistik Deskriptif mencakup perhitungan ukuran pemusatan (seperti Mean dan Median) dan ukuran penyebaran (seperti Standard Deviation dan IQR) untuk meringkas data. Sementara itu, Eksplorasi Data Visual, menggunakan alat seperti Histogram, Box Plot, dan Bar Chart, berfungsi untuk mengidentifikasi pola distribusi, mendeteksi keberadaan pencilan (outlier), serta memvisualisasikan proporsi kategorisasi status gizi (IMT) mahasiswa.

Key Findings

Berdasarkan analisis statistik deskriptif dan eksplorasi data visual:

- Rata-rata IMT: Rata-rata Indeks Massa Tubuh (IMT) mahasiswa adalah 21.942, namun nilai tengahnya (Median) yang lebih akurat menggambarkan pusat data adalah 20.519.
- Distribusi Data: Data IMT menunjukkan distribusi yang sangat miring ke kanan (highly positive skewed) dengan nilai Skewness 9.951, yang disebabkan oleh keberadaan pencilan ekstrim.
- Deteksi Pencilan (Outlier): Terdapat pencilan (outlier) IMT yang signifikan, beberapa di antaranya mencapai nilai yang sangat tinggi (hingga 156.107), yang harus diperhatikan dalam interpretasi data.
- Status Gizi Dominan: Mayoritas mahasiswa memiliki Status Gizi Normal (62.9%). Kategori Kurus (24.5%) merupakan kategori non-normal terbesar.

1. Pendahuluan

1.1 Latar Belakang

Tujuan utama dari tugas besar ini adalah menunjukkan kemampuan mahasiswa dalam mengimplementasikan metode statistik pada kasus dunia nyata. Penelitian ini menggunakan data Tinggi Badan dan Berat Badan mahasiswa untuk menghitung IMT dan mengkategorikan status gizi.

1.2 Tujuan

- Menunjukkan kemampuan mahasiswa dalam pemahaman metode statistik. khususnya ukuran pemusatan dan penyebaran.
- Menunjukkan kemampuan mahasiswa dalam membersihkan data dan menghitung variabel turunan (IMT).
- Menunjukkan kemampuan mahasiswa dalam meng-implementasi metode statistik pada kasus dunia asli.
- Menunjukkan kemampuan mahasiswa dalam penggunaan bahasa pemrograman R untuk analisis data.

2. Metodologi

Metode analisis utama yang digunakan adalah Statistik Deskriptif dan Eksplorasi Data Visual untuk memahami distribusi dan karakteristik data Tinggi Badan, Berat Badan, dan IMT.

2.1 Pertanyaan Penelitian

- Bagaimana ukuran pemusatan dan penyebaran data untuk Tinggi Badan, Berat Badan, dan IMT mahasiswa?
- Bagaimana distribusi data IMT, dan apakah terdapat pencilan (outlier)?
- Bagaimana distribusi persentase status gizi mahasiswa berdasarkan kategori IMT?

2.2 Desain Penelitian

Tabel berikut akan menyimpulkan dan menjelaskan metode - metode untuk menjawab setiap pertanyaan

Pertanyaan Penelitian	Metode yang digunakan untuk menjawab
Ukuran pemusatan dan penyebaran?	Mean, Median, Standar Deviasi, IQR
Distribusi dan pencilan data?	Histogram, Skewness, Box Plot
Persentase Status Gizi?	Kategorisasi IMT, Bar Chart

2.3 Proses Pembersihan dan Standardisasi Data (RStudio)

Proses pembersihan data dilakukan secara menyeluruh menggunakan RStudio. Tahapan pertama dimulai dengan mengimpor data, kemudian merestrukturisasi nama-nama kolom yang sebelumnya tidak beraturan agar memiliki format yang konsisten dan mudah dibaca. Selanjutnya, kualitas data diperbaiki dengan mengonversi seluruh kolom ke tipe teks terlebih dahulu untuk mencegah kesalahan tipe data, serta mengganti nilai kosong seperti "" atau "-" menjadi nilai hilang (NA). Pada tahap standardisasi numerik, karakter non-angka pada kolom numerik dihapus dan nilai hilang pada variabel tertentu, seperti IPK, diisi menggunakan metode imputasi berbasis median. Untuk variabel kategorik, penulisan kategori distandardisasi agar setiap kategori memiliki bentuk yang seragam. Setelah data bersih, variabel turunan berupa Indeks Massa Tubuh (IMT) dihitung menggunakan rumus $IMT = \text{Berat Badan (kg)} / (\text{Tinggi Badan dalam meter})^2$. Nilai IMT tersebut kemudian diklasifikasikan ke dalam kategori Kurus, Normal, Kelebihan Berat, dan Obesitas berdasarkan standar WHO Asia Pasifik.

2.4 Coding untuk Statistik Deskriptif dan Eksplorasi

Berikut adalah *script* R yang digunakan untuk membersihkan data antropometri, menghitung IMT, dan menghasilkan statistik deskriptif serta visualisasi eksplorasi data:

```
{r}
# --- 1. MEMUAT LIBRARY YANG DIBUTUHKAN ---
# install.packages(c("dplyr", "ggplot2", "e1071")) # Hapus tanda '#' jika belum
# terinstal
library(dplyr)
library(ggplot2)
library(e1071)

# --- 2. MEMUAT DAN MEMBERSIHKAN DATA SERTA MENGHITUNG IMT ---
df_data <- read.csv("data_ads_bersih.csv")

# Daftar variabel antropometri
vars_antropometri <- c("Tinggi_Badan", "Berat_Badan")

# Pembersihan data dan perhitungan IMT
df_clean_imt <- df_data %>%
  select(all_of(vars_antropometri)) %>%
  mutate_all(~as.numeric(as.character(.))) %>%
  na.omit() %>%
  # HITUNG IMT (Berat_Badan / (Tinggi_Badan dalam meter)^2)
  mutate(IMT = Berat_Badan / ((Tinggi_Badan / 100)^2)) %>%
  # Kategorisasi IMT (Menggunakan standar Asia Pasifik WHO)
  mutate(Status_Gizi = case_when(
    IMT < 18.5 ~ "Kurus (Underweight)",
    IMT >= 18.5 & IMT < 25 ~ "Normal",
    IMT >= 25 & IMT < 30 ~ "Kelebihan Berat (Overweight)",
    IMT >= 30 ~ "Obesitas",
    TRUE ~ "Tidak Terdefinisi"
  ))

# Daftar semua variabel numerik yang akan dianalisis (TB, BB, IMT)
vars_to_analyze <- c("Tinggi_Badan", "Berat_Badan", "IMT")
```

```

# -----
# FUNGSI UNTUK MENGHITUNG STATISTIK DESKRIPTIF LENGKAP
# -----

hitung_deskriptif_lengkap <- function(x) {
  x_clean <- na.omit(x)

  if (length(x_clean) < 2) return(NULL)

  data.frame(
    N = length(x_clean),
    Mean = mean(x_clean),
    Median = median(x_clean),
    Std_Dev = sd(x_clean),
    Min = min(x_clean),
    Max = max(x_clean),
    IQR = IQR(x_clean),
    Skewness = skewness(x_clean)
  ) %>% round(3)
}

# -----
# 3. STATISTIK DESKRIPTIF (PEMUSATAN, PENYEBARAN, SKEWNESS)
# -----

cat("## 📊 3. Statistik Deskriptif: TB, BB, dan IMT\n")
cat("-----\n")

deskriptif_final <- NULL

for (var in vars_to_analyze) {
  hasil <- hitung_deskriptif_lengkap(df_clean_imt[[var]])
  if (!is.null(hasil)) {
    hasil$Variabel <- gsub("_", " ", var)
    deskriptif_final <- rbind(deskriptif_final, hasil)
  }
}

print(deskriptif_final)

```

```

cat("## 4. Visualisasi Pola Fisik dan Status Gizi\n")
cat("-----\n")

# A. Visualisasi Distribusi IMT (Histogram)
print(
  ggplot(df_clean_imt, aes(x = IMT)) +
  geom_histogram(aes(y = after_stat(density)), bins = 20, fill = "darkgreen", color = "white", alpha = 0.7) +
  labs(title = "Histogram Distribusi Indeks Massa Tubuh (IMT)",
        x = "IMT",
        y = "Kepadatan") +
  theme_minimal()
)

# B. Visualisasi Deteksi Pencilan IMT (Box Plot)
print(
  ggplot(df_clean_imt, aes(y = IMT)) +
  geom_boxplot(fill = "lightgreen", outlier.colour = "red", outlier.shape = 8) +
  labs(title = "Box Plot (Deteksi Pencilan) Indeks Massa Tubuh",
        y = "IMT") +
  theme_minimal()
)

# C. Visualisasi Status Gizi (Bar Chart)
print(
  df_clean_imt %>%
  group_by(Status_Gizi) %>%
  summarise(N = n()) %>%
  mutate(Persentase = N / sum(N)) %>%
  ggplot(aes(x = reorder(Status_Gizi, Persentase), y = Persentase, fill = Status_Gizi)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = scales::percent(Persentase)), vjust = -0.5) +
  labs(title = "Distribusi Status Gizi Mahasiswa (Berdasarkan IMT)",
        x = "Status Gizi (Kategori IMT)",
        y = "Persentase") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
)

```

3. Hasil dan Diskusi

3.1 Statistik Deskriptif: Tinggi Badan, Berat Badan, dan IMT

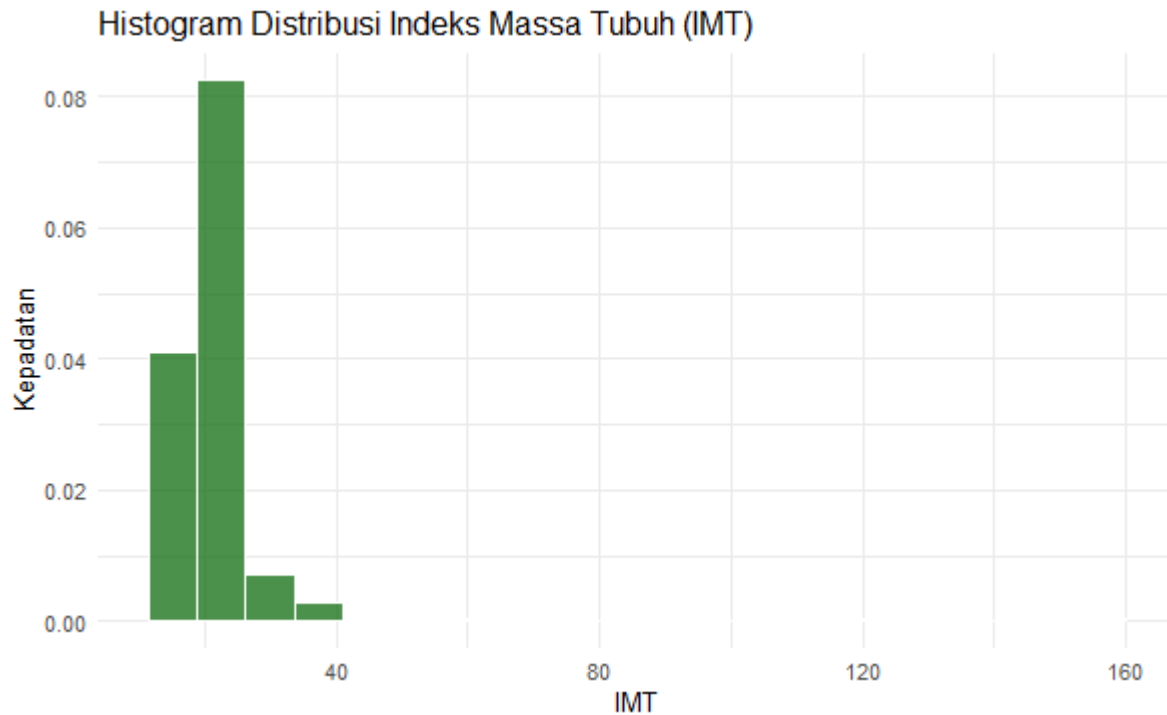
Analisis dilakukan pada total 379 observasi.

Variabel	N	Mean	Median	Std_Dev	Min	Max	IQR	Skewness
Tinggi Badan	379	161.570	161.000	11.576	63.000	192.000	14.000	-3.065
Berat Badan	379	56.317	53.000	14.053	33.000	160.000	15.000	2.047
IMT	379	21.942	20.519	9.671	15.111	156.107	4.379	9.951

Perbedaan besar antara Mean (21.942) dan Median (20.519) IMT, serta nilai Skewness yang sangat tinggi (9.951), mengindikasikan bahwa data IMT sangat miring ke kanan dan dipengaruhi oleh nilai maksimum yang ekstrem (156.107).

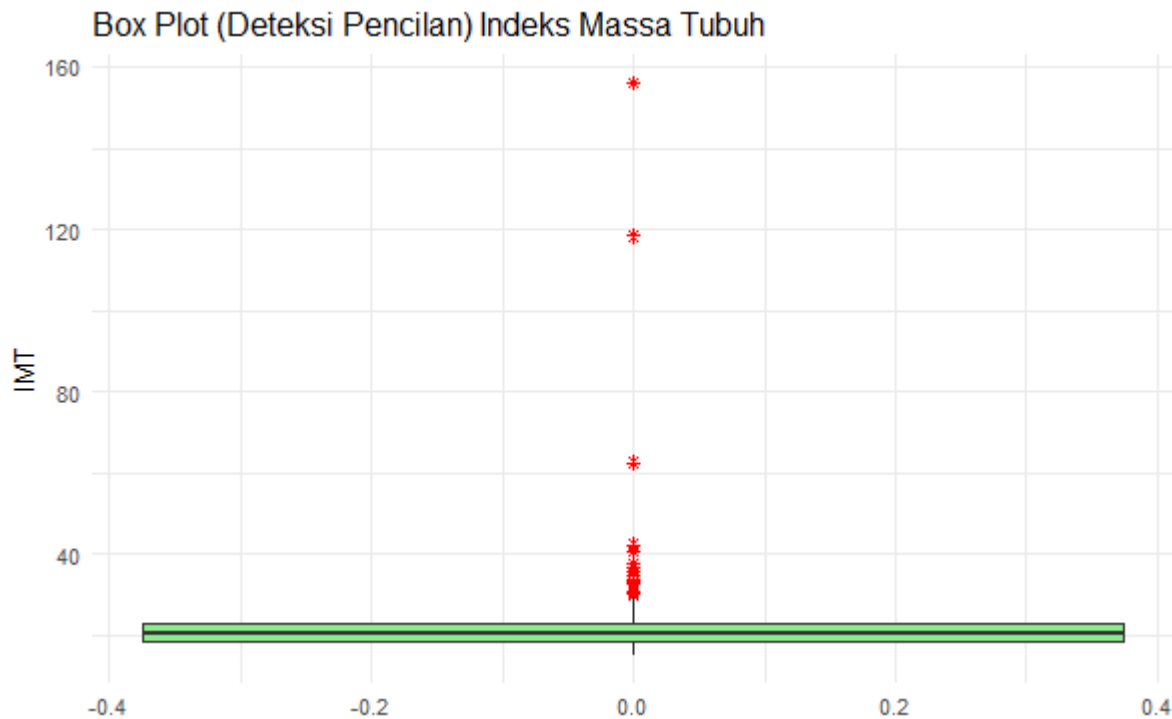
3.2 Eksplorasi Data Visual

3.2.1 Distribusi Indeks Massa Tubuh (IMT)



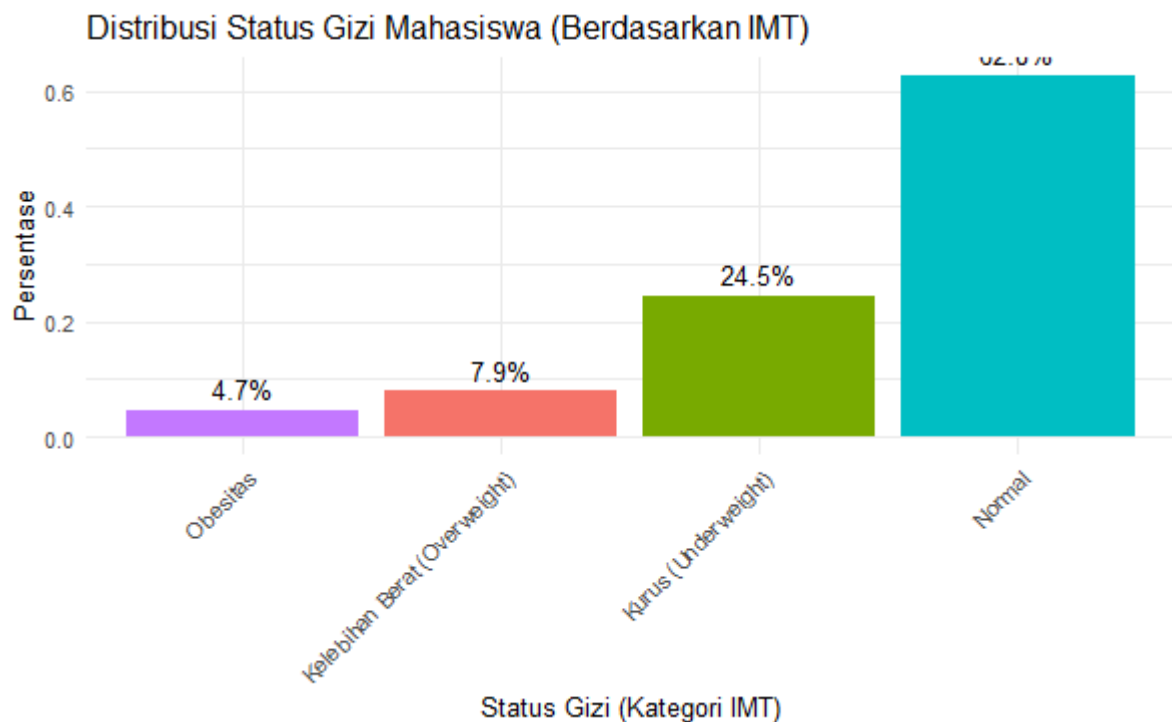
Histogram menunjukkan bahwa kepadatan data IMT terpusat pada rentang yang rendah (Normal), dengan ekor yang sangat panjang ke kanan, secara visual mengonfirmasi positive skew ekstrem yang terdeteksi secara numerik.

3.2.2 Deteksi Pencilan (*Outlier*) IMT



Box Plot secara dramatis menunjukkan keberadaan pencilan ekstrem (outlier, bintang merah) yang jauh di atas kuartil. Nilai-nilai ekstrem ini, yang tertinggi mencapai 156.107, menjelaskan mengapa nilai Mean IMT tertarik jauh dari Median dan menyebabkan Skewness yang sangat tinggi.

3.2.3 Distribusi Status Gizi Mahasiswa



Bar Chart Status Gizi menunjukkan:

- Kategori Normal adalah yang paling dominan dengan 62.9%.
- Kategori Kurus (Underweight) menempati persentase terbesar kedua (24.5%).
- Gabungan kategori Kelebihan Berat (Overweight) dan Obesitas adalah 12.6%.

4. Kesimpulan

Temuan Kunci (Key Findings)

Berdasarkan analisis statistik deskriptif dan eksplorasi data visual:

1. IMT Tengah: Nilai tengah (Median) IMT mahasiswa adalah 20.519. Nilai ini lebih representatif daripada Mean (21.942) karena nilai Mean dipengaruhi kuat oleh pencilan.
2. Pencilan Ekstrem: Data IMT memiliki pencilan (outlier) ekstrem yang signifikan, dengan nilai tertinggi mencapai 156.107. Hal ini menyebabkan distribusi sangat miring ke kanan (highly positive skewed).
3. Status Gizi Dominan: Mayoritas mahasiswa memiliki Status Gizi Normal (62.9%). Kategori Kurus (24.5%) merupakan kategori non-normal terbesar.

Referensi

1. Populis / Fanruan. Analisis Data Eksplorasi (EDA): Arti, Manfaat, dan Contohnya. Diakses dari [Analisis Data Eksplorasi \(EDA\): Arti, Manfaat, dan Contohnya](#).
2. World Health Organization (WHO). (2000). The Asia-Pacific Perspective: Redefining Obesity and its Treatment. Western Pacific Region.