Kelompok 6:

1. *Kevin Antoni Junior (* 123450109)
2. *Muhammad Daffa Arrafi (*124450120)
3. *Nazlah Auliya (*124450054)
4. *Riska Erlis Dayu (*124450022)

1. IMPORT DATA

```{r}
data <- read.csv("C:/Users/asus/Downloads/Dataset Tugas Besar ADS 2025 - Karakteristik Mahasiswa .csv")
```

2. CEK STRUKTUR DATA

Penjelasan:
Baris ini digunakan untuk mengambil (mengimpor) dataset dari file CSV yang ada di komputer. Setelah dijalankan, seluruh isi file akan masuk ke dalam sebuah objek bernama data sehingga bisa diproses di RStudio.

2. CEK STRUKTUR DATA

```{r}
str(data)
summary(data)
head(data)
```

Description: df [6 × 20]

	NIM <int>	Program.Studi.. <chr>	IPK.Terakhir.. <chr>	Jenis.Kelamin <chr>	Tinggi.Badan.. <chr>	Berat.Badan.. <chr>	Pendidikan.terakhir <chr>
1	1	Sains Data	4. 0	Laki-laki	160	50	SMA
2	2	Matematika	3.8	Laki-laki	169	57	SMA
3	3	Sains Data	3.4	Perempuan	150	39	SMA
4	4	Sains Data	3.86	Laki-laki	170	65	SMA
5	5	Sains Data	3.97	Laki-laki	170	63	SMA
6	6	Sains Data	3.35	Perempuan	154	40	SMA

6 rows | 1-8 of 20 columns

Penjelasan:

str(data) Menampilkan bentuk dan jenis tiap kolom dalam dataset. Dari output terlihat ada 20 kolom berisi data seperti NIM, Program Studi, IPK, Jenis Kelamin, tinggi, berat, dan lainnya. Ini berguna untuk memastikan tipe datanya sudah benar.

summary(data) Memberikan ringkasan statistik untuk tiap kolom, seperti nilai minimum–maksimum untuk angka dan jumlah kategori untuk data teks. Tujuannya agar kita cepat memahami gambaran umum isi dataset.

head(data) Menampilkan 6 baris pertama sebagai contoh isi data. Output menunjukkan beberapa mahasiswa dengan informasi lengkap, sehingga kita bisa cek apakah data berhasil terbaca dengan baik

3. PEMBERSIHAN DATA DAN UBAH NAMA COLOM AGAR LEBIH MUDAH DAN KOVERIS TIPE DATA YANG HARUSNYA NUMERIK

```r
library(dplyr)
library(stringr)
library(readr)
#mengganti nama colom
nama_baru <- c(
  "nim", "prodi", "ipk", "gender", "tinggi", "berat", "pendidikan",
  "jam_belajar", "beasiswa", "asal_daerah", "pekerjaan", "internet",
  "organisasi", "uang_saku", "tempat_tinggal", "jarak", "ayah", "ibu",
  "pendapatan", "keluarga"
)

  colnames(data) <- nama_baru
# FUNGSI PEMBERSIH ANGKA
clean_mixed_numeric <- function(x) {
  x <- as.character(x)
  # Hapus teks 'jam', ganti koma jadi titik, hapus spasi
  x <- str_remove_all(str_to_lower(x), "jam|hours")
  x <- str_replace_all(x, ",", ".")
  x <- str_remove_all(x, "\\s")
  # Ambil angka saja
  num <- str_extract(x, "[0-9]+(\\.[0-9]+)?")
  return(as.numeric(num))
}

#BERSIHKAN DATA
data_clean <- data %>%
  mutate(
    # Bersihkan kolom angka
    ipk = clean_mixed_numeric(ipk),
    jam_belajar = clean_mixed_numeric(jam_belajar),
    tinggi = as.numeric(tinggi),
    berat = as.numeric(berat),

    # Rapikan teks (Title Case & Hapus Spasi) pada semua kolom teks
    across(where(is.character), ~ str_to_title(str_trim(.))),

    # Ubah "-" atau kosong jadi NA
    across(where(is.character), ~ na_if(., "-")),
    across(where(is.character), ~ na_if(., "")),

    # Standardisasi Nama Prodi
    prodi = str_replace(prodi, "Sains Data.*", "Sains Data")
  ) %>%
  distinct() # Hapus duplikat
data_clean <- data_clean %>%
  filter(ipk >= 0 & ipk <= 4)

#Konversi tipe data yang seharusnya numerik
num_vars <- c("ipk", "tinggi", "berat", "jam_belajar")

data_clean[num_vars] <- lapply(data_clean[num_vars], as.numeric)
# menghapus data yang tidak ada ipk
data_clean <- data_clean[!is.na(data_clean$ipk), ]
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
i In argument: `tinggi = as.numeric(tinggi)`.
Caused by warning:
! NAs introduced by coercion
i Run dplyr::last_dplyr_warnings() to see the 1 remaining warning.
```

Penjelasan:

Kode tersebut digunakan untuk membersihkan data agar rapi dan siap dipakai untuk analisis. Pertama, nama-nama kolom diganti agar lebih pendek dan mudah dibaca. Setelah itu dibuat fungsi khusus bernama clean_mixed_numeric() untuk membersihkan kolom angka yang berisi

campuran teks, seperti “5 jam” atau “3 hours”, lalu mengambil angka saja dan mengubahnya menjadi format numerik. Pada bagian mutate(), beberapa kolom seperti ipk, jam_belajar, tinggi, dan berat dibersihkan dari teks, tanda baca, dan spasi, lalu diubah menjadi angka. Semua kolom berbentuk teks kemudian dirapikan, misalnya huruf dibuat kapital di awal kata dan spasi berlebih dihapus. Tanda “-” atau sel kosong diubah menjadi NA agar mudah dibersihkan. Nama prodi juga distandarkan supaya seragam menjadi “Sains Data”. Setelah itu, baris duplikat dihapus dan data disaring agar IPK hanya berada di rentang 0 sampai 4. Terakhir, kolom-kolom yang seharusnya numerik dikonversi ulang menjadi angka dan baris yang memiliki NA pada IPK dibuang.

Uji Heteroskedastisitas

```{r}
library(lmtest)
skeditas <- bptest(model)
skeditas
P_value <- skeditas$p.value
alpha <- 0.05
if(p_value < alpha){
  cat("ada heteroskedastisitas")
}else{
  cat("tiadk ada heteroskedastisitas")
}
```

```
        studentized Breusch-Pagan test

data:  model
BP = 1.5392, df = 3, p-value = 0.6732

ada heteroskedastisitas
```

Penjelasan:

Uji ini melihat apakah penyebaran kesalahan (residual) stabil atau berubah-ubah. Idealnya residual harus memiliki penyebaran yang sama. Kalau p-value > 0.05, berarti tidak ada heteroskedastisitas (aman).

p-value 0.6732 (>0.05) → Secara statistik hasilnya aman

```
Uji  multikolinearitas
```{r}
library(car)
mlt <- vif(model)
mlt
if(all(mlt < 10)){
   cat("Tidak ada multikolinearitas")
}else{
   cat("ada multikolinearitas")
}

```
```
Warning: package 'car' was built under R version 4.5.2
Loading required package: carData
Warning: package 'carData' was built under R version 4.5.2
jam_belajar      tinggi       berat
   1.008718    1.256639    1.261870
Tidak ada multikolinearitas
```

Penjelasan:

Uji ini dipakai untuk mengecek apakah variabel bebas (misalnya jam belajar, tinggi, berat) saling “mirip” atau saling mempengaruhi secara berlebihan.Kalau angkanya di atas 0.1, berarti aman dan tidak ada masalah multikolinearitas.

Artinya: setiap variabel benar-benar memberikan informasi yang berbeda, jadi bisa dipakai bareng-bareng dalam model regresi. Nilai semua variabel > 1, jadi aman, tidak ada multikolinearitas.
```

Uji normalitas

```r
library(MASS)
stdres <- stdres(model)
stdr <- data.frame(stdres)
sapiro <- shapiro.test(stdres)
sapiro
p_value <- sapiro$p.value
alpha <- 0.05
if(p_value < alpha){
  cat("Data tidak terdistribusi normal")
}else{
  cat("Data terdistribusi normal")
}
```

```
        Shapiro-Wilk normality test

data:  stdres
W = 0.71457, p-value < 2.2e-16

Data tidak terdistribusi normal
```

Penjelasan:

Uji ini dipakai untuk melihat apakah data residual (selisih antara nilai prediksi dan nilai asli) bentuknya mengikuti pola normal atau tidak.

Kalau p-value < 0.05, berarti data tidak normal. p-value < 2.2e-16 (sangat kecil), artinya data residual tidak berdistribusi normal.

```r
head(data_clean)
```

Description: df [6 × 20]

	nim <int>	prodi <chr>	ipk <dbl>	gender <chr>	tinggi <dbl>	berat <dbl>	pendidikan <chr>	jam_belajar <dbl>	beasiswa <chr>
1	1	Sains Data	4.00	Laki-Laki	160	50	Sma	3.0	Tidak
2	2	Matematika	3.80	Laki-Laki	169	57	Sma	48.0	Tidak
3	3	Sains Data	3.40	Perempuan	150	39	Sma	17.5	Tidak
4	4	Sains Data	3.86	Laki-Laki	170	65	Sma	4.0	Tidak
5	5	Sains Data	3.97	Laki-Laki	170	63	Sma	30.0	Tidak
6	6	Sains Data	3.35	Perempuan	154	40	Sma	3.0	Tidak

6 rows | 1-10 of 20 columns

Penjelasan:

Output head(data_clean) menampilkan enam baris pertama dari data yang sudah dibersihkan. Terlihat bahwa kolom-kolom seperti nim, prodi, ipk, gender, tinggi, berat, pendidikan, jam_belajar, dan beasiswa sudah rapi dan berada dalam format yang benar. Kolom angka seperti ipk, tinggi, berat, dan jam_belajar sudah berubah menjadi numerik tanpa campuran teks, sehingga siap dipakai untuk analisis statistik. Kolom teks seperti prodi, gender, dan pendidikan

sudah ditata ulang dengan huruf yang konsisten dan lebih mudah dibaca. Data ini menunjukkan bahwa proses pembersihan berjalan dengan baik: tidak ada karakter aneh, penulisan sudah rapi, dan tipe data sudah sesuai dengan kebutuhan analisis. Hasilnya, dataset sekarang jauh lebih bersih dan siap dipakai untuk tahap analisis berikutnya.