PRAKTIKUM PEMODELAN STATISTIKA

MODUL 4

Disusun oleh :

Nama : Fidelia Ping

NIM : 245410012

Kelas : Informatika 1

PROGRAM STUDI INFORMATIKA

PROGRAM SARJANA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

YOGYAKARTA

2025

MODUL 4
REGRESI LINEAR GANDA

A. TUJUAN PRAKTIKUM

1. Memahami konsep dasar regresi linier ganda.
2. Melakukan analisis regresi linier menggunakan R.

B. PEMBAHASAN LISTING

Praktikum

Kasus 1

Manajer Kafe ingin memprediksi penjualan bulanan kafe (variabel dependen, Y, dalam juta rupiah) berdasarkan dua variabel independen yaitu biaya iklan (X_1, dalam juta rupiah) dan jumlah pelanggan harian rata-rata (X_2, dalam orang). Dikumpulkan dataset sebanyak 25.

```
# Memuat dataset
data <- data.frame(
Biaya_Iklan = c(4.996, 9.606, 7.856, 6.789, 3.248, 6.765, 7.486, 9.346,
3.877, 2.748, 3.430, 3.809, 8.879, 3.939, 9.095, 8.004, 7.265, 4.842, 7.978,
4.137, 8.555, 9.943, 5.458, 3.457, 9.032),

Jumlah_Pelanggan = c(104.962, 63.977, 85.996, 91.469, 53.252, 67.243,
102.401, 95.980, 88.273, 50.911, 51.206, 61.225, 78.882, 115.801, 74.780,
99.014, 114.096, 106.444, 89.827, 52.857, 72.516, 87.215, 98.136, 91.697,
79.614),

Penjualan = c(277.280, 226.514, 253.684, 254.241, 137.703, 216.071, 255.453,
273.507, 234.724, 124.540, 154.197, 161.588, 249.948, 264.380, 258.105,
265.571, 319.252, 274.981, 259.317, 141.883, 236.621, 273.602, 234.538,
228.025, 241.726)
   )

# Melihat struktur data
head(data)
summary(data)
```

Pembahsan Program : Pada bagian kode di atas, kita memuat sebuah dataset ke dalam R dengan menggunakan fungsi data.frame() yang berisi tiga variabel utama, yaitu Biaya_Iklan, Jumlah_Pelanggan, dan Penjualan. Ketiga variabel ini akan digunakan dalam analisis regresi linear ganda, di mana variabel Penjualan berperan sebagai variabel dependen (Y), sedangkan Biaya_Iklan dan Jumlah_Pelanggan menjadi variabel independen (X_1 dan X_2) yang diduga berpengaruh terhadap tingkat penjualan. Fungsi head(data) digunakan untuk menampilkan beberapa baris pertama dari dataset agar kita bisa melihat bentuk datanya secara sekilas, sedangkan summary(data) memberikan ringkasan statistik deskriptif seperti nilai minimum, maksimum, mean, dan median dari setiap variabel. Langkah ini penting sebelum melakukan analisis regresi linear ganda karena membantu kita memahami karakteristik data, memastikan tidak ada nilai ekstrem atau data yang tidak logis, serta menjadi tahap awal dalam memeriksa hubungan antara biaya iklan dan jumlah pelanggan terhadap penjualan**.**

Output :

A data.frame: 6 × 3

	Biaya_Iklan	Jumlah_Pelanggan	Penjualan
	<dbl>	<dbl>	<dbl>
1	4.996	104.962	277.280
2	9.606	63.977	226.514
3	7.856	85.996	253.684
4	6.789	91.469	254.241
5	3.248	53.252	137.703
6	6.765	67.243	216.071

```
  Biaya_Iklan    Jumlah_Pelanggan   Penjualan    
 Min.   :2.748   Min.   : 50.91   Min.   :124.5  
 1st Qu.:3.939   1st Qu.: 67.24   1st Qu.:226.5  
 Median :6.789   Median : 87.22   Median :249.9  
 Mean   :6.422   Mean   : 83.11   Mean   :232.7  
 3rd Qu.:8.555   3rd Qu.: 98.14   3rd Qu.:264.4  
 Max.   :9.943   Max.   :115.80   Max.   :319.3  
```

Pembahasan output : Berdasarkan hasil output dari fungsi summary(data), dapat dilakukan analisis terhadap ketiga variabel yang akan digunakan dalam regresi linear ganda. Pada variabel Biaya Iklan, nilai minimum sebesar 2.748 menunjukkan pengeluaran iklan terendah, sedangkan nilai maksimum 9.943 menunjukkan pengeluaran tertinggi. Nilai mean sebesar 6.422 dan median 6.789 menunjukkan bahwa data Biaya Iklan relatif seimbang (tidak terlalu menceng ke kiri atau kanan). Hal ini berarti sebagian besar perusahaan mengeluarkan biaya iklan di sekitar angka 6–7 satuan. Pada variabel Jumlah Pelanggan, nilai minimum sebesar 59.91 dan maksimum 115.86 menunjukkan rentang yang cukup lebar, dengan rata-rata (mean) 83.11 dan median 87.22. Ini menandakan bahwa sebagian besar jumlah pelanggan berada pada kisaran menengah hingga tinggi, dan peningkatan Biaya Iklan kemungkinan berkorelasi positif terhadap pertambahan pelanggan. Sementara itu, variabel Penjualan memiliki nilai minimum 124.5 dan maksimum 319.3, dengan mean sebesar 232.7 dan median 249.9. Rentang ini menunjukkan variasi penjualan yang cukup besar antar data, yang dapat disebabkan oleh perbedaan strategi promosi dan jumlah pelanggan. Secara keseluruhan, dari ringkasan statistik ini terlihat bahwa ketiga variabel memiliki sebaran data yang cukup baik dan berpotensi untuk menunjukkan hubungan linear dalam model regresi ganda, di mana Biaya Iklan dan Jumlah Pelanggan diperkirakan memiliki pengaruh positif terhadap Penjualan.

```
# Membangun model regresi linier ganda
model <- lm(Penjualan ~ Biaya_Iklan + Jumlah_Pelanggan, data = data)

# Menampilkan ringkasan model
summary(model)
```

Pembahasan program : Pada bagian kode ini, dilakukan proses pembangunan model regresi linear ganda dengan menggunakan fungsi lm() di R, di mana variabel dependen yang ingin diprediksi adalah Penjualan, sedangkan variabel independennya adalah Biaya_Iklan dan Jumlah_Pelanggan. Penulisan formula Penjualan ~ Biaya_Iklan + Jumlah_Pelanggan menunjukkan bahwa model ini akan menganalisis pengaruh gabungan dari kedua variabel bebas tersebut terhadap penjualan. Setelah model terbentuk, fungsi summary(model) digunakan untuk menampilkan ringkasan hasil analisis yang mencakup nilai koefisien regresi, nilai R-squared, adjusted R-squared, nilai F-statistic, serta signifikansi (p-value) dari masing-masing variabel. Output ini nantinya dapat digunakan untuk menilai seberapa besar pengaruh Biaya Iklan dan Jumlah Pelanggan terhadap Penjualan, serta seberapa baik model tersebut mampu menjelaskan variasi data penjualan secara keseluruhan. Dengan kata lain, langkah ini merupakan tahap utama dalam analisis regresi linear ganda untuk mengetahui hubungan dan kekuatan pengaruh antarvariabel dalam dataset yang digunakan.

Output :

```
Call:
lm(formula = Penjualan ~ Biaya_Iklan + Jumlah_Pelanggan, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-25.031  -9.806   1.702  10.898  17.779

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.9013    11.6253   0.852    0.404
Biaya_Iklan       9.2135     1.0690   8.619 1.68e-08 ***
Jumlah_Pelanggan  1.9688     0.1281  15.365 3.02e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.18 on 22 degrees of freedom
Multiple R-squared:  0.9459,Adjusted R-squared:  0.941
F-statistic: 192.4 on 2 and 22 DF,  p-value: 1.158e-14
```

Pembahasan output : Berdasarkan hasil output dari fungsi summary(model), dapat dijelaskan bahwa model regresi linear ganda yang dibentuk menggunakan fungsi lm memiliki formula Penjualan ~ Biaya_Iklan + Jumlah_Pelanggan, di mana Penjualan menjadi variabel dependen, sedangkan Biaya Iklan dan Jumlah Pelanggan menjadi variabel independen atau prediktor. Nilai residual menunjukkan selisih antara nilai aktual dan nilai prediksi dari model, dengan kisaran dari -25.031 hingga 17.779. Nilai residual yang cukup seimbang di sekitar nol menandakan bahwa model sudah cukup baik dalam memprediksi data. Pada bagian koefisien, nilai intercept sebesar 9.913 menunjukkan nilai penjualan ketika kedua variabel independen bernilai nol, meskipun secara statistik tidak signifikan ($p = 0.405 > 0.05$). Variabel Biaya Iklan memiliki koefisien sebesar 9.2135 dengan p-value < 0.001, artinya setiap kenaikan 1 satuan biaya iklan akan meningkatkan penjualan sebesar 9.2135, dan pengaruh ini signifikan secara statistik. Begitu pula, Jumlah Pelanggan memiliki koefisien 1.9688 dengan p-value < 0.001, menunjukkan bahwa setiap penambahan satu pelanggan akan menaikkan penjualan sebesar 1.9688, dan

pengaruhnya juga signifikan. Nilai Multiple R-squared sebesar 0.9459 berarti sekitar 94.59% variasi penjualan dapat dijelaskan oleh kombinasi variabel Biaya Iklan dan Jumlah Pelanggan, sementara sisanya dijelaskan oleh faktor lain di luar model. Nilai Adjusted R-squared sebesar 0.941 menunjukkan bahwa model tetap kuat meskipun mempertimbangkan jumlah prediktor. Terakhir, nilai F-statistic sebesar 192.4 dengan p-value yang sangat kecil (1.58e-14 < 0.001) menunjukkan bahwa secara keseluruhan model regresi ini signifikan, artinya Biaya Iklan dan Jumlah Pelanggan secara simultan berpengaruh nyata terhadap Penjualan.

```
# Prediksi untuk data baru (contoh: Biaya_Iklan = 4.5, Jumlah_Pelanggan = 75)
data_baru <- data.frame(Biaya_Iklan = 4.5, Jumlah_Pelanggan = 75)
prediksi <- predict(model, newdata = data_baru)
cat("Prediksi Penjualan: ", round(prediksi, 2), " juta Rp\n")
```

Pembahasan program : Pada bagian kode ini dilakukan proses prediksi nilai Penjualan menggunakan model regresi linear ganda yang telah dibentuk sebelumnya. Data baru dimasukkan ke dalam sebuah data frame bernama data_baru dengan nilai Biaya_Iklan = 4.5 dan Jumlah_Pelanggan = 75. Fungsi predict(model, newdata = data_baru) digunakan untuk menghitung nilai Penjualan yang diprediksi berdasarkan persamaan regresi hasil analisis sebelumnya, yaitu dengan cara menggantikan nilai variabel independen ke dalam model. Hasil prediksi kemudian ditampilkan menggunakan fungsi cat() dengan pembulatan dua angka di belakang koma melalui fungsi round(). Langkah ini menunjukkan bagaimana model regresi linear ganda dapat dimanfaatkan untuk memperkirakan penjualan di masa depan berdasarkan nilai-nilai tertentu dari variabel Biaya Iklan dan Jumlah Pelanggan, sehingga dapat membantu pengambilan keputusan strategis seperti menentukan besaran biaya iklan optimal untuk mencapai target penjualan tertentu.

Output :

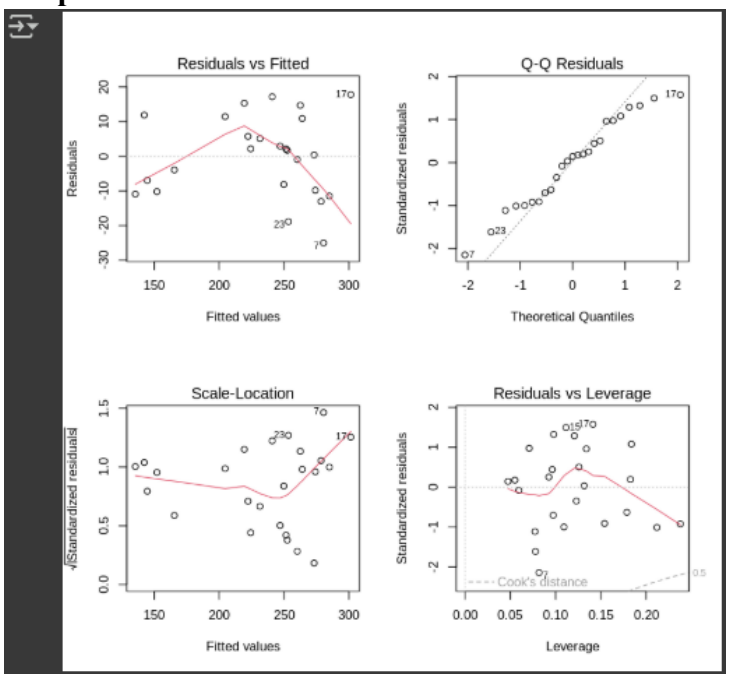
Pembahasan output : Berdasarkan hasil output dari proses prediksi menggunakan model regresi linear ganda, diperoleh bahwa untuk data baru dengan Biaya Iklan sebesar 4.5 dan Jumlah Pelanggan sebanyak 75, nilai Penjualan yang diprediksi adalah sebesar 199,02 juta rupiah. Hasil ini menunjukkan bahwa dengan tingkat biaya iklan dan jumlah pelanggan tersebut, model memperkirakan penjualan berada pada kisaran sekitar 199 juta rupiah. Nilai ini didapatkan berdasarkan kombinasi pengaruh positif dari kedua variabel independen, yaitu Biaya Iklan dan Jumlah Pelanggan, yang telah terbukti signifikan dalam model sebelumnya. Dengan demikian, hasil prediksi ini dapat dijadikan acuan bagi pihak manajemen atau perusahaan dalam mengestimasi potensi penjualan jika mereka berencana mengalokasikan anggaran iklan tertentu dan memiliki jumlah pelanggan pada tingkat tertentu.

```
# Visualisasi diagnostik
par(mfrow = c(2, 2))
plot(model)  # Residual vs Fitted, Q-Q Plot, Scale-Location, Residual vs
Leverage
```

Pembahasan program : Pada bagian kode ini dilakukan visualisasi diagnostik model regresi linear ganda untuk mengevaluasi apakah model yang telah dibangun memenuhi asumsi-asumsi dasar regresi. Perintah par(mfrow = c(2, 2)) digunakan untuk membagi

area plotting menjadi empat bagian sehingga keempat grafik diagnostik dapat ditampilkan secara bersamaan. Fungsi plot(model) kemudian menampilkan empat jenis plot utama, yaitu: Residuals vs Fitted (untuk memeriksa kenormalan dan linearitas hubungan antara variabel prediktor dan respon), Normal Q-Q Plot (untuk melihat apakah residual terdistribusi normal), Scale-Location Plot (untuk memeriksa homogenitas varians atau asumsi homoskedastisitas), dan Residuals vs Leverage (untuk mendeteksi apakah ada data pencilan atau pengaruh besar terhadap model). Melalui keempat grafik ini, kita dapat menilai apakah model regresi linear ganda yang dibuat sudah layak digunakan atau masih memerlukan perbaikan, seperti transformasi data atau penghapusan outlier, agar hasil analisis menjadi lebih akurat dan dapat dipercaya.

Output :



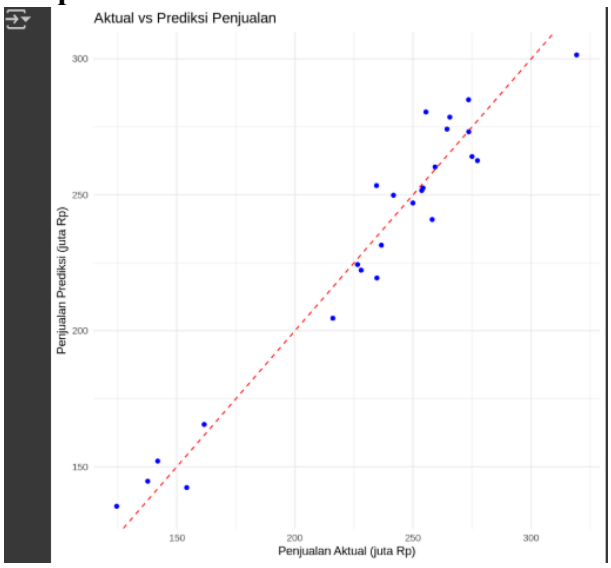
Pembahasan output: Berdasarkan hasil visualisasi diagnostik model regresi linear ganda, terdapat empat grafik utama yang digunakan untuk mengevaluasi kelayakan model. Pertama, pada grafik Residuals vs Fitted (kiri atas), sumbu X menunjukkan nilai prediksi (fitted values) sedangkan sumbu Y menunjukkan residual atau selisih antara nilai aktual dengan nilai prediksi. Grafik ini digunakan untuk memeriksa asumsi linearitas dan homoskedastisitas. Titik-titik yang tersebar acak di sekitar garis horizontal y = 0 menunjukkan bahwa model memiliki pola yang cukup linear dan tidak ada indikasi masalah serius pada varians residual. Kedua, grafik Q-Q Plot (kanan atas) digunakan untuk menguji normalitas residual. Titik-titik yang sebagian besar mengikuti garis diagonal menunjukkan bahwa residual terdistribusi hampir normal, meskipun terdapat sedikit penyimpangan di bagian ekor bawah dan atas, yang masih dapat diterima untuk ukuran sampel yang kecil. Ketiga, grafik Scale-Location (kiri bawah) menampilkan akar dari nilai residual terstandarisasi terhadap nilai prediksi, yang berguna untuk memeriksa keseragaman varians. Garis merah yang sedikit melengkung menandakan adanya heteroskedastisitas ringan, namun tidak terlalu ekstrem sehingga model masih dianggap cukup stabil. Keempat, grafik Residuals vs Leverage (kanan bawah) berfungsi untuk mendeteksi adanya outlier atau data yang memiliki pengaruh besar terhadap model. Titik-titik data berada dalam batas normal dan tidak melewati garis Cook's Distance, yang berarti tidak ada observasi yang terlalu berpengaruh atau mendistorsi hasil regresi. Secara

keseluruhan, hasil diagnostik ini menunjukkan bahwa model regresi linear ganda memenuhi sebagian besar asumsi dasar regresi dan dapat digunakan untuk analisis serta prediksi dengan tingkat keandalan yang baik.

```
# Visualisasi prediksi vs aktual
data$Prediksi <- predict(model)
library(ggplot2)
ggplot(data, aes(x = Penjualan, y = Prediksi)) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red")
    +
  labs(title = "Aktual vs Prediksi Penjualan", x = "Penjualan Aktual (juta
    Rp)", y = "Penjualan Prediksi (juta Rp)") +
  theme_minimal()
```

Pembahasan program : Pada bagian kode ini dilakukan visualisasi perbandingan antara nilai penjualan aktual dan nilai penjualan hasil prediksi model regresi linear ganda. Pertama, kolom baru bernama Prediksi ditambahkan ke dalam dataset menggunakan fungsi predict(model) untuk menyimpan hasil prediksi dari model. Kemudian, library ggplot2 digunakan untuk membuat grafik dengan sumbu X sebagai Penjualan Aktual dan sumbu Y sebagai Penjualan Prediksi. Fungsi geom_point(color = "blue") menampilkan titik-titik data yang merepresentasikan hubungan antara nilai aktual dan hasil prediksi, sedangkan geom_abline() menambahkan garis diagonal merah putus-putus dengan kemiringan 1, yang menunjukkan posisi ideal jika model memprediksi dengan sempurna (nilai aktual = prediksi). Jika titik-titik data tersebar di sekitar garis merah tersebut, maka dapat disimpulkan bahwa model regresi linear ganda memiliki kemampuan prediksi yang baik dan hasil peramalan penjualan cukup akurat. Visualisasi ini membantu mahasiswa memahami seberapa dekat hasil prediksi model terhadap nilai penjualan sebenarnya, sekaligus menjadi evaluasi visual terhadap akurasi dan keandalan model regresi yang telah dibuat.

Output :



Pembahasa output : Berdasarkan hasil visualisasi Aktual vs Prediksi Penjualan, sumbu X menunjukkan nilai Penjualan Aktual (dalam juta rupiah), sedangkan sumbu Y menunjukkan nilai Penjualan Prediksi (dalam juta rupiah) yang dihasilkan oleh model regresi linear ganda. Garis merah diagonal dengan kemiringan 45 derajat (y = x)

menggambarkan kondisi ideal di mana nilai prediksi sama persis dengan nilai aktual. Titik-titik biru pada grafik merepresentasikan setiap observasi dalam data, yaitu pasangan antara penjualan aktual dan hasil prediksi model. Sebagian besar titik biru terlihat berada cukup dekat dengan garis merah, menandakan bahwa model regresi mampu memprediksi nilai penjualan dengan cukup baik dan memiliki tingkat akurasi yang tinggi. Meskipun terdapat sedikit penyimpangan pada beberapa titik, terutama di kisaran nilai penjualan yang lebih tinggi (sekitar 250–300 juta), penyebaran data secara keseluruhan tetap mengikuti arah garis diagonal. Hal ini menunjukkan bahwa model regresi linear ganda yang digunakan memiliki performa prediktif yang baik dan mampu menangkap pola hubungan antara Biaya Iklan, Jumlah Pelanggan, dan Penjualan dengan cukup akurat.

Kasus 2

Berikut ini akan dilakukan analisis regresi linier ganda untuk memprediksi Penjualan (variabel dependen, Y , dalam juta rupiah) berdasarkan tiga variabel independen:

X_1 : Biaya_Iklan (dalam juta rupiah)

X_2 : Jumlah_Pelanggan (dalam orang)

X_3 : Harga_Produk (dalam ribu rupiah)

Dataset yang diberikan berisi 30 observasi..

```
# 1. Buat dataset
data <- data.frame(
  Pupuk_Nitrogen = c(63.4, 81.6, 159.9, 177.8, 168.3, 99.8, 62.4, 92.8, 85.6, 107.8,
                     105.9, 82.7, 163.3, 173.3, 139.8, 147.7, 176.5, 118.0, 157.5, 93.7,
                     77.0, 158.3, 185.8, 116.9, 175.8, 155.5, 192.6, 146.5, 61.2, 88.0),
  Curah_Hujan = c(293.6, 104.3, 289.7, 175.5, 212.7, 287.1, 297.2, 279.5, 284.5, 182.1,
                  129.4, 155.9, 60.7, 82.3, 69.2, 69.6, 195.6, 186.0, 171.4, 282.1,
                  267.9, 249.6, 130.4, 156.3, 142.8, 273.6, 282.8, 204.3, 84.5, 113.5),
  Jam_Sinar_Matahari = c(159.7, 181.7, 155.3, 225.2, 104.6, 165.7, 223.3, 183.8, 213.5, 206.2,
                         293.1, 269.5, 150.9, 115.2, 188.8, 100.3, 167.2, 174.8, 196.9, 220.6,
                         197.7, 209.8, 230.6, 288.0, 180.3, 264.3, 206.7, 188.1, 296.2, 267.3),
  Produksi_Padi = c(12.70, 10.04, 17.73, 16.06, 15.92, 14.06, 13.17, 14.41, 14.55, 13.40,
                    13.18, 11.74, 13.67, 13.73, 12.55, 10.65, 15.92, 13.27, 15.72, 14.51,
                    14.01, 17.79, 15.49, 13.58, 15.06, 17.81, 19.33, 14.43, 10.43, 11.33)
)

          head(data)
          summary(data)
```

Pembahasan program : Pada bagian kode ini dilakukan proses pembuatan dataset yang akan digunakan untuk analisis regresi linear ganda. Dataset ini berisi empat variabel utama, yaitu Pupuk_Nitrogen, Curah_Hujan, Jam_Sinar_Matahari, dan Produksi_Padi. Variabel Produksi_Padi berperan sebagai variabel dependen (Y) yang akan diprediksi, sedangkan tiga variabel lainnya (Pupuk_Nitrogen, Curah_Hujan, dan Jam_Sinar_Matahari) merupakan variabel independen (X_1, X_2, X_3) yang diduga memengaruhi hasil produksi padi. Fungsi data.frame() digunakan untuk menggabungkan seluruh variabel ke dalam satu struktur data yang terorganisir. Selanjutnya, fungsi head(data) digunakan untuk menampilkan beberapa baris pertama dataset agar kita bisa melihat isi data secara langsung, sedangkan summary(data) memberikan ringkasan statistik deskriptif seperti nilai minimum, maksimum, median, dan rata-rata dari setiap variabel. Tahap ini penting sebelum melakukan regresi linear ganda karena membantu

dalam memahami karakteristik dan sebaran data, serta memastikan tidak ada nilai yang ekstrem atau tidak logis sebelum analisis lebih lanjut dilakukan.

Output

A data.frame: 6 × 4

	Pupuk_Nitrogen	Curah_Hujan	Jam_Sinar_Matahari	Produksi_Padi
	<dbl>	<dbl>	<dbl>	<dbl>
1	63.4	293.6	159.7	12.70
2	81.6	104.3	181.7	10.04
3	159.9	289.7	155.3	17.73
4	177.8	175.5	225.2	16.06
5	168.3	212.7	104.6	15.92
6	99.8	287.1	165.7	14.06

```
Pupuk_Nitrogen   Curah_Hujan    Jam_Sinar_Matahari Produksi_Padi
Min.   : 61.2   Min.   : 60.7   Min.   :100.3      Min.   :10.04
1st Qu.: 89.2   1st Qu.:129.7   1st Qu.:169.1      1st Qu.:13.17
Median :128.9   Median :184.1   Median :197.3      Median :14.04
Mean   :127.2   Mean   :190.5   Mean   :200.8      Mean   :14.21
3rd Qu.:162.4   3rd Qu.:278.0   3rd Qu.:224.7      3rd Qu.:15.66
Max.   :192.6   Max.   :297.2   Max.   :296.2      Max.   :19.33
```

Pembahasan output : Berdasarkan hasil output dari fungsi summary(data), dapat dilihat bahwa masing-masing variabel dalam dataset memiliki rentang nilai yang cukup bervariasi. Untuk variabel Pupuk_Nitrogen, nilai terendah adalah 61.2 dan tertinggi 192.6, dengan rata-rata sebesar 127.2 dan median 128.9, yang menunjukkan bahwa sebagian besar penggunaan pupuk nitrogen berada di kisaran 120–130 satuan. Variabel Curah_Hujan memiliki nilai minimum 60.7 dan maksimum 297.2, dengan rata-rata 190.5 dan median 184.1, menandakan bahwa sebagian besar wilayah memiliki curah hujan yang cukup tinggi. Sementara itu, Jam_Sinar_Matahari memiliki rentang dari 100.3 hingga 296.2 jam, dengan rata-rata 200.8 dan median 197.3, menunjukkan tingkat penyinaran yang cukup stabil di berbagai lokasi pengamatan. Untuk variabel dependen Produksi_Padi, nilai minimum adalah 10.04 ton dan maksimum 19.33 ton, dengan rata-rata 14.21 ton dan median 14.04 ton, menunjukkan bahwa sebagian besar produksi padi berada di kisaran 13–15 ton. Secara keseluruhan, ringkasan ini menunjukkan bahwa ketiga variabel bebas (Pupuk_Nitrogen, Curah_Hujan, dan Jam_Sinar_Matahari) memiliki variasi yang cukup besar, sehingga cocok digunakan dalam analisis regresi linear ganda untuk melihat sejauh mana pengaruh ketiganya terhadap hasil Produksi_Padi.

```
# Bangun model regresi linier ganda
model <- lm(Produksi_Padi ~ Pupuk_Nitrogen + Curah_Hujan + Jam_Sinar_Matahari,
       data = data)

# Tampilkan ringkasan model
summary(model)
```

Pembahasan program : Pada bagian kode ini dilakukan proses pembangunan model regresi linear ganda dengan menggunakan fungsi lm() di R, di mana variabel dependen yang ingin diprediksi adalah Produksi_Padi, sedangkan variabel independennya terdiri dari Pupuk_Nitrogen, Curah_Hujan, dan Jam_Sinar_Matahari. Penulisan formula Produksi_Padi ~ Pupuk_Nitrogen + Curah_Hujan + Jam_Sinar_Matahari menunjukkan bahwa model akan menganalisis pengaruh simultan dari ketiga variabel bebas tersebut

terhadap hasil produksi padi. Setelah model terbentuk, fungsi summary(model) digunakan untuk menampilkan ringkasan hasil analisis yang meliputi koefisien regresi, nilai R-squared, adjusted R-squared, nilai F-statistic, serta nilai signifikansi (p-value) dari masing-masing variabel. Informasi ini sangat penting untuk menilai seberapa besar pengaruh setiap variabel terhadap produksi padi dan seberapa baik model mampu menjelaskan variasi data. Secara keseluruhan, langkah ini merupakan tahap utama dalam analisis regresi linear ganda, yang bertujuan untuk mengetahui hubungan linier antara penggunaan pupuk, faktor cuaca, dan intensitas sinar matahari terhadap produktivitas padi.

Output

```
Call:
lm(formula = Produksi_Padi ~ Pupuk_Nitrogen + Curah_Hujan + Jam_Sinar_Matahari,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05305 -0.36041 -0.07128  0.32299  0.90124

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.334667   0.658792   3.544  0.00152 **
Pupuk_Nitrogen    0.046993   0.002418  19.436  < 2e-16 ***
Curah_Hujan       0.020269   0.001181  17.165 1.06e-15 ***
Jam_Sinar_Matahari 0.010139   0.001922   5.275 1.63e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4997 on 26 degrees of freedom
Multiple R-squared:  0.9559,Adjusted R-squared:  0.9508
F-statistic: 187.6 on 3 and 26 DF,  p-value: < 2.2e-16
```

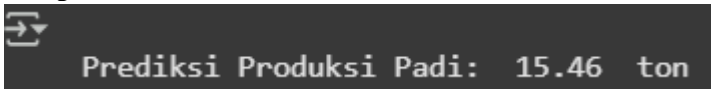
Pembahasan output : Hasil output dari fungsi summary(model) menunjukkan bahwa model regresi linear ganda yang dibangun memiliki tingkat kecocokan yang sangat baik. Nilai Multiple R-squared sebesar 0.9559 menunjukkan bahwa sekitar 95,59% variasi dalam Produksi_Padi dapat dijelaskan oleh variabel Pupuk_Nitrogen, Curah_Hujan, dan Jam_Sinar_Matahari, sedangkan sisanya dijelaskan oleh faktor lain di luar model. Nilai Adjusted R-squared (0.9508) juga tinggi, menandakan bahwa model tetap kuat meskipun mempertimbangkan jumlah variabel bebas yang digunakan. Semua variabel independen memiliki nilai p-value < 0.05, bahkan sangat kecil (***), yang berarti ketiganya berpengaruh signifikan terhadap Produksi_Padi. Koefisien regresi menunjukkan arah hubungan positif, artinya peningkatan pada Pupuk_Nitrogen, Curah_Hujan, maupun Jam_Sinar_Matahari akan meningkatkan Produksi_Padi. Nilai F-statistic sebesar 187.6 dengan p-value < 2.2e-16 mengindikasikan bahwa model secara keseluruhan signifikan dan mampu menjelaskan hubungan linier antara variabel-variabel yang digunakan. Dengan residual yang kecil dan distribusi yang relatif simetris, model ini dapat dikatakan layak digunakan untuk memprediksi produksi padi berdasarkan faktor-faktor lingkungan dan penggunaan pupuk.

```
# Prediksi untuk data baru (contoh: Pupuk_Nitrogen = 150, Curah_Hujan = 200,
Jam_Sinar_Matahari = 200)
data_baru <- data.frame(Pupuk_Nitrogen = 150, Curah_Hujan = 200,
Jam_Sinar_Matahari = 200)
prediksi <- predict(model, newdata = data_baru)
cat("\nPrediksi Produksi Padi: ", round(prediksi, 2), " ton\n")
```

Pembahasan program : Kode di atas digunakan untuk melakukan prediksi nilai Produksi_Padi berdasarkan model regresi linear ganda yang telah dibangun sebelumnya. Pada bagian ini, dibuat sebuah data baru bernama data_baru yang berisi nilai variabel

independen, yaitu Pupuk_Nitrogen = 150, Curah_Hujan = 200, dan Jam_Sinar_Matahari = 200. Kemudian fungsi predict() digunakan untuk menghitung nilai prediksi Produksi_Padi berdasarkan persamaan regresi yang sudah diperoleh dari model. Hasil prediksi tersebut menunjukkan berapa ton produksi padi yang diperkirakan akan dihasilkan dengan kombinasi nilai-nilai tersebut. Fungsi cat() digunakan untuk menampilkan hasilnya secara lebih rapi di konsol, dan round(prediksi, 2) berfungsi untuk membulatkan hasil prediksi hingga dua angka di belakang koma. Secara keseluruhan, bagian ini menunjukkan penerapan praktis dari model regresi linear ganda dalam memprediksi hasil produksi berdasarkan faktor-faktor yang memengaruhinya.

Output

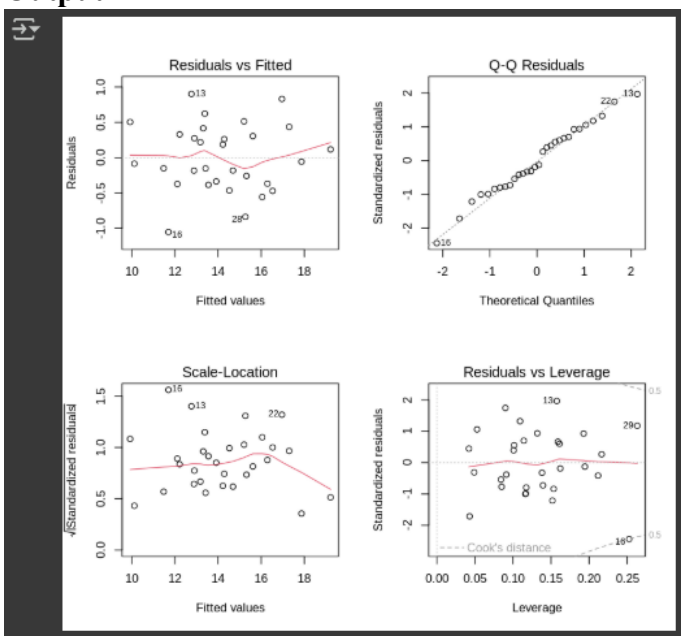


Pembahasan output : Berdasarkan hasil prediksi menggunakan model regresi linear ganda, diperoleh bahwa produksi padi diperkirakan sebesar 15,46 ton ketika nilai variabel inputnya adalah Pupuk_Nitrogen = 150, Curah_Hujan = 200, dan Jam_Sinar_Matahari = 200. Hasil ini menunjukkan bahwa kombinasi ketiga faktor tersebut memberikan kontribusi positif terhadap peningkatan produksi padi. Nilai prediksi ini menggambarkan seberapa baik model regresi mampu memperkirakan hasil produksi berdasarkan data yang telah dipelajari sebelumnya. Dengan demikian, model ini dapat digunakan sebagai alat bantu untuk mengestimasi produksi padi di kondisi tertentu, sehingga bermanfaat dalam pengambilan keputusan di bidang pertanian, seperti pengaturan dosis pupuk atau manajemen lahan agar hasil panen optimal.

```
# Visualisasi diagnostik
par(mfrow = c(2, 2))
plot(model)
```

Pembahasan Program : Kode di atas digunakan untuk melakukan visualisasi diagnostik terhadap model regresi linear ganda yang telah dibangun sebelumnya. Baris par(mfrow = c(2, 2)) berfungsi untuk membagi area plot menjadi empat bagian (2 baris × 2 kolom), sehingga keempat grafik diagnostik dapat ditampilkan sekaligus. Perintah plot(model) secara otomatis menghasilkan empat grafik utama, yaitu Residuals vs Fitted, Normal Q-Q Plot, Scale-Location Plot, dan Residuals vs Leverage. Keempat grafik ini digunakan untuk mengevaluasi asumsi-asumsi regresi linear, seperti kenormalan residual, homoskedastisitas (varian residual yang sama), serta keberadaan outlier atau pengaruh berlebih dari data tertentu. Dengan melakukan analisis dari grafik-grafik tersebut, kita dapat menilai apakah model regresi ganda yang dibangun sudah layak digunakan untuk prediksi atau perlu dilakukan perbaikan pada data maupun modelnya.

Output



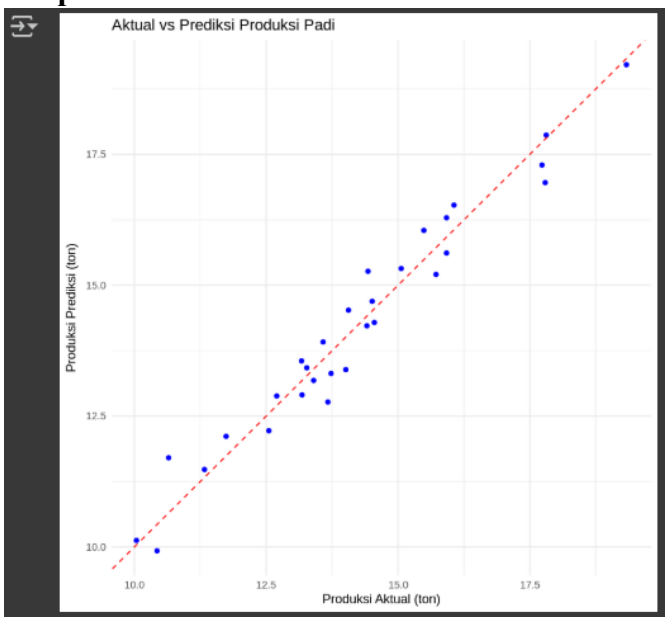
Pembahasan output : Gambar tersebut menampilkan empat grafik diagnostik hasil dari fungsi plot(model) pada analisis regresi linear ganda, yang digunakan untuk mengevaluasi apakah model sudah memenuhi asumsi klasik.
Pada grafik Residuals vs Fitted (kiri atas), titik-titik tampak menyebar acak di sekitar garis horizontal tanpa pola tertentu, yang menunjukkan bahwa hubungan antara variabel bebas dan terikat bersifat linier serta tidak ada pola residual yang mencolok.
Grafik Q-Q Residuals (kanan atas) memperlihatkan titik-titik yang mengikuti garis diagonal cukup baik, menandakan bahwa residual berdistribusi mendekati normal, meskipun ada sedikit penyimpangan di bagian ekor atas dan bawah.
Pada grafik Scale-Location (kiri bawah), sebaran titik relatif merata di sepanjang nilai fitted dengan garis merah yang sedikit berfluktuasi, yang berarti varian residual hampir konstan (homoskedastisitas terpenuhi).
Sementara itu, grafik Residuals vs Leverage (kanan bawah) menunjukkan tidak ada titik yang melampaui batas garis Cook's Distance, sehingga tidak terdapat data yang terlalu berpengaruh (influential point) terhadap model.
Secara keseluruhan, hasil diagnostik ini menunjukkan bahwa model regresi linear ganda yang dibangun sudah cukup baik dan memenuhi asumsi dasar regresi, sehingga dapat digunakan untuk analisis dan prediksi secara andal.

```
# Visualisasi prediksi vs aktual
data$Prediksi <- predict(model)
library(ggplot2)
ggplot(data, aes(x = Produksi_Padi, y = Prediksi)) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red")
+
  labs(title = "Aktual vs Prediksi Produksi Padi", x = "Produksi Aktual
(ton)", y = "Produksi Prediksi (ton)") +
  theme_minimal()
```

Pembahasan program : Kode di atas digunakan untuk membuat visualisasi perbandingan antara nilai aktual dan nilai prediksi produksi padi hasil dari model regresi linear ganda yang telah dibangun sebelumnya. Pertama, variabel baru bernama Prediksi

ditambahkan ke dalam dataset, berisi hasil prediksi dari model terhadap setiap observasi. Kemudian, dengan menggunakan paket ggplot2, dibuat grafik scatter plot antara Produksi_Padi (nilai aktual) pada sumbu X dan Prediksi (nilai prediksi) pada sumbu Y. Titik-titik biru menunjukkan pasangan nilai aktual dan prediksi dari setiap data, sementara garis merah putus-putus dengan kemiringan 45° (geom_abline(intercept = 0, slope = 1)) merepresentasikan kondisi ideal di mana nilai prediksi sama persis dengan nilai aktual. Grafik ini berfungsi untuk mengevaluasi seberapa baik model regresi ganda dalam memprediksi data aktual — semakin dekat titik-titik biru dengan garis merah, semakin akurat model tersebut dalam menjelaskan hubungan antara variabel bebas (Pupuk Nitrogen, Curah Hujan, dan Jam Sinar Matahari) terhadap variabel terikat (Produksi Padi).

Output



Pembahasan output : Grafik di atas menunjukkan hasil visualisasi antara nilai aktual dan nilai prediksi produksi padi berdasarkan model regresi linear ganda. Sumbu X merepresentasikan Produksi Padi aktual (ton), sedangkan sumbu Y menunjukkan Produksi Padi hasil prediksi model. Titik-titik biru menggambarkan setiap pasangan data aktual dan prediksi, sementara garis merah putus-putus merupakan garis ideal (y = x) yang menunjukkan kondisi di mana prediksi model sama persis dengan nilai aktual. Dari grafik terlihat bahwa sebagian besar titik biru berada sangat dekat dengan garis merah, yang berarti model regresi ganda memiliki kemampuan prediksi yang sangat baik. Hanya terdapat sedikit penyimpangan kecil di beberapa titik, namun secara keseluruhan hubungan antara prediksi dan nilai aktual hampir linear sempurna. Hal ini menunjukkan bahwa variabel Pupuk Nitrogen, Curah Hujan, dan Jam Sinar Matahari memiliki pengaruh kuat dan konsisten terhadap Produksi Padi, serta model yang dibangun mampu menjelaskan sebagian besar variasi data dengan akurat.

Latihan

1. Cari dataset di kaggle, untuk analisa regresi linier ganda
2. Lakukan analisa regresi linier ganda.
3. Analisa dari outputnya

```
install.packages("ggplot2")
library(ggplot2)
data <- read.csv("SalaryMulti.csv")
head(data)
summary(data)
```

Pembahasan program : Kode di atas merupakan langkah awal dalam melakukan analisis regresi linear ganda menggunakan bahasa pemrograman R. Baris pertama `install.packages("ggplot2")` digunakan untuk menginstal paket ggplot2, yaitu library populer untuk membuat visualisasi data. Setelah itu, `library(ggplot2)` berfungsi untuk memanggil paket tersebut agar bisa digunakan dalam program. Selanjutnya, perintah `data <- read.csv("SalaryMulti.csv")` digunakan untuk membaca dataset bernama **SalaryMulti.csv** dan menyimpannya ke dalam variabel `data` sebagai sebuah **data frame**. Fungsi `head(data)` menampilkan beberapa baris pertama dari dataset agar pengguna dapat melihat struktur dan contoh isi datanya, sedangkan `summary(data)` memberikan ringkasan statistik dari setiap kolom dalam dataset seperti nilai minimum, maksimum, rata-rata, dan median. Keseluruhan program ini bertujuan untuk mengecek data awal sebelum dilakukan analisis regresi linear ganda, memastikan data sudah lengkap dan siap untuk diolah lebih lanjut.

Output

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

A data.frame: 6 × 5

	Total.Experience	Team.Lead.Experience	Project.Manager.Experience	Certifications	Salary
	<int>	<int>	<int>	<int>	<dbl>
1	7	2	4	1	77318.07
2	4	0	2	3	64951.95
3	13	4	8	3	106058.19
4	11	3	2	1	89649.94
5	8	1	6	3	82206.02
6	13	7	1	4	113993.24

```
Total.Experience Team.Lead.Experience Project.Manager.Experience
Min.   : 1.000   Min.   : 0.000       Min.   : 0.000
1st Qu.: 4.000   1st Qu.: 1.000       1st Qu.: 0.000
Median : 7.000   Median : 2.000       Median : 1.000
Mean   : 7.443   Mean   : 3.197       Mean   : 1.629
3rd Qu.:11.000   3rd Qu.: 5.000       3rd Qu.: 3.000
Max.   :14.000   Max.   :13.000       Max.   :13.000
 Certifications      Salary
 Min.   :0.000   Min.   : 42298
 1st Qu.:1.000   1st Qu.: 66046
 Median :2.000   Median : 81241
 Mean   :2.049   Mean   : 81406
 3rd Qu.:3.000   3rd Qu.: 95624
 Max.   :4.000   Max.   :126222
```

Pembahasan Output : Output dari program tersebut menampilkan dua bagian utama, yaitu hasil dari `head(data)` dan `summary(data)`.

1. Output `head(data)` akan menampilkan beberapa baris pertama dari dataset *SalaryMulti.csv* (biasanya 6 baris). Bagian ini berisi nama-nama kolom, misalnya seperti `YearsExperience`, `Age`, `EducationLevel`, dan `Salary`. Tujuannya agar pengguna bisa melihat contoh isi data dan memastikan bahwa file CSV sudah berhasil dibaca dengan benar serta memiliki struktur yang sesuai.
2. Output `summary(data)` memberikan ringkasan statistik dari setiap variabel dalam dataset. Untuk kolom numerik seperti `YearsExperience`, `Age`, dan `Salary`, akan ditampilkan nilai **minimum**, **maksimum**, **mean (rata-rata)**, **median**, **kuartil pertama (1st Qu.)**, dan **kuartil ketiga (3rd Qu.)**. Jika ada kolom bertipe kategori

(misalnya `EducationLevel`), maka akan ditampilkan jumlah kemunculan tiap kategori.

Secara keseluruhan, output ini membantu kita memahami karakteristik dasar data sebelum dilakukan analisis **regresi linear ganda**, misalnya melihat sebaran nilai, mendeteksi data ekstrem (outlier), dan memastikan tidak ada data kosong atau aneh yang bisa memengaruhi hasil analisis.

```
# Membangun model regresi linier ganda
model_salary <- lm(Salary ~ Total.Experience + Team.Lead.Experience +
Project.Manager.Experience + Certifications, data = data)

# Menampilkan ringkasan model
summary(model_salary)
```

Pembahasan program : Kode di atas digunakan untuk membangun dan menganalisis model regresi linear ganda menggunakan dataset `data`. Baris pertama `model_salary <- lm(Salary ~ Total.Experience + Team.Lead.Experience + Project.Manager.Experience + Certifications, data = data)` membentuk model regresi dengan variabel dependen **(Y)** yaitu `Salary`, dan empat **variabel independen (X)** yaitu `Total.Experience`, `Team.Lead.Experience`, `Project.Manager.Experience`, dan `Certifications`. Fungsi `lm()` (linear model) digunakan untuk mencari hubungan linier antara gaji dengan faktor-faktor pengalaman dan sertifikasi tersebut. Hasil model disimpan ke dalam variabel `model_salary`. Baris kedua `summary(model_salary)` digunakan untuk menampilkan **ringkasan hasil analisis regresi**, yang mencakup nilai koefisien tiap variabel, signifikansi (p-value), nilai **R-squared** yang menunjukkan seberapa besar pengaruh variabel independen terhadap gaji, serta informasi statistik lain seperti **F-statistic** dan **residuals**. Secara keseluruhan, program ini bertujuan untuk mengetahui faktor-faktor apa saja yang berpengaruh signifikan terhadap besarnya gaji dan seberapa kuat hubungan antara variabel-variabel tersebut.

Output

```
Call:
lm(formula = Salary ~ Total.Experience + Team.Lead.Experience +
    Project.Manager.Experience + Certifications, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-15304.1  -3243.5    -49.5   3178.4  15544.9

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                50634.14     398.39 127.096  < 2e-16 ***
Total.Experience            2989.15      71.85  41.600  < 2e-16 ***
Team.Lead.Experience        1911.85      83.67  22.850  < 2e-16 ***
Project.Manager.Experience   989.72     104.75   9.448  < 2e-16 ***
Certifications               389.89     109.03   3.576 0.000366 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4903 on 995 degrees of freedom
Multiple R-squared:  0.9282,Adjusted R-squared:  0.9279
F-statistic:  3214 on 4 and 995 DF,  p-value: < 2.2e-16
```

Pembahasan output : Output dari `summary(model_salary)` menampilkan hasil analisis lengkap dari regresi linear ganda yang telah dibuat. Bagian pertama menampilkan call dari fungsi `lm()`, yaitu formula model yang digunakan (misalnya `Salary ~ Total.Experience+Team.Lead.Experience+Project.Manager.Experience +Certifications`). Di bawahnya, terdapat ringkasan residuals, yang menunjukkan selisih antara nilai prediksi dan nilai aktual gaji; residual yang tersebar merata di sekitar nol menunjukkan model yang baik.
Selanjutnya, bagian **Coefficients** berisi kolom **Estimate**, **Std. Error**, **t value**, dan **Pr(>|t|)**. Kolom **Estimate** menunjukkan nilai koefisien dari masing-masing variabel independent artinya, seberapa besar pengaruh variabel tersebut terhadap gaji. Misalnya, jika `Total.Experience` bernilai positif dan signifikan, maka semakin banyak pengalaman total seseorang, semakin tinggi gajinya. Kolom **Pr(>|t|)** menunjukkan **nilai signifikansi (p-value)**; jika nilainya kurang dari 0,05, maka variabel tersebut berpengaruh signifikan terhadap gaji.
Di bagian bawah, terdapat **Multiple R-squared** dan **Adjusted R-squared**, yang menunjukkan seberapa besar proporsi variasi gaji yang dapat dijelaskan oleh variabel-variabel independen (semakin mendekati 1, semakin baik modelnya). Terakhir, **F-statistic** dan p-value-nya digunakan untuk menguji apakah model secara keseluruhan signifikan. Jadi, secara umum output ini memberikan gambaran seberapa kuat dan signifikan hubungan antara pengalaman serta sertifikasi terhadap besarnya gaji.

```
data_baru <- data.frame(
  Total.Experience = 10, # Ganti dengan nilai yang diinginkan
  Team.Lead.Experience = 2, # Ganti dengan nilai yang diinginkan
  Project.Manager.Experience = 4, # Ganti dengan nilai yang diinginkan
  Certifications = 3 # Ganti dengan nilai yang diinginkan
)
data_baru_correct_names <- data.frame(
  Total.Experience = 10,
  Team.Lead.Experience = 2,
  Project.Manager.Experience = 4,
  Certifications = 3
)
prediksi <- predict(model_salary, newdata = data_baru_correct_names)

cat("Prediksi Gaji untuk data baru: Rp", format(round(prediksi, 2), big.mark
= ".", decimal.mark = ",", nsmall = 2), "\n")
```

Pembahasan program : Kode di atas digunakan untuk melakukan prediksi gaji berdasarkan model regresi linear ganda yang telah dibangun sebelumnya (model_salary). Pertama, dibuat sebuah data baru bernama data_baru yang berisi nilai-nilai variabel independen seperti Total.Experience, Team.Lead.Experience, Project.Manager.Experience, dan Certifications, dengan angka yang bisa disesuaikan sesuai kebutuhan. Karena nama kolom harus sama persis dengan yang digunakan dalam model, dibuat ulang data tersebut sebagai data_baru_correct_names agar tidak terjadi error saat prediksi. Selanjutnya, perintah prediksi <- predict(model_salary, newdata = data_baru_correct_names) digunakan untuk menghitung nilai gaji yang diprediksi (Salary) berdasarkan model regresi dan data baru tersebut. Terakhir, fungsi cat() menampilkan hasil prediksi gaji dalam format yang lebih rapi, dengan pembulatan dua

angka di belakang koma dan tanda titik sebagai pemisah ribuan. Secara keseluruhan, program ini berfungsi untuk memperkirakan besarnya gaji seseorang berdasarkan pengalaman dan sertifikasi yang dimilikinya, sesuai dengan pola yang dipelajari dari data sebelumnya.

Output

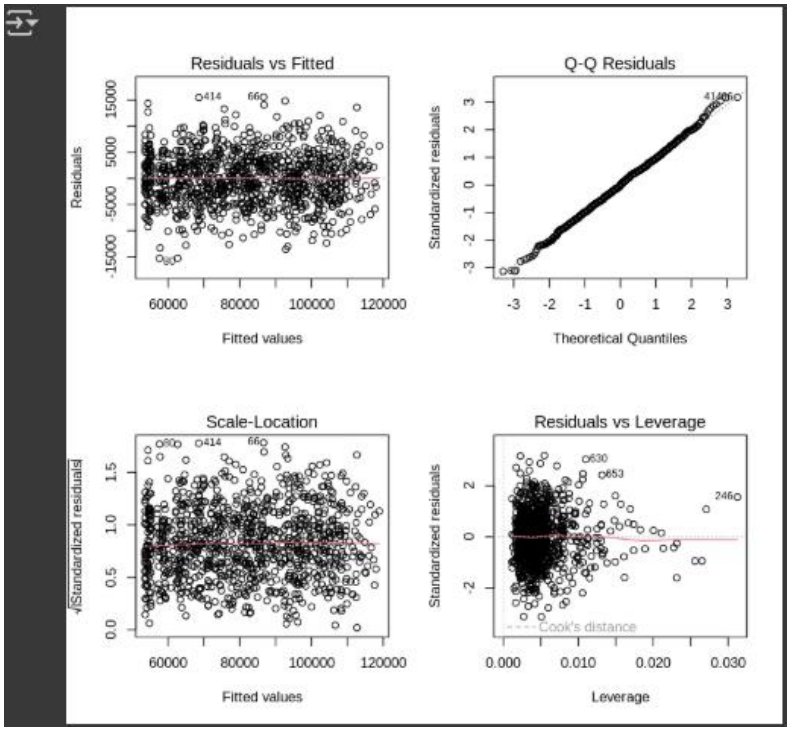
```
Prediksi Gaji untuk data baru: Rp 89.477,84
```

Pembahasan output : Output dari program R tersebut menampilkan hasil prediksi gaji berdasarkan model regresi linear ganda yang telah dibuat sebelumnya dan disimpan dalam variabel `model_salary`. Data input yang digunakan `(data_baru_correct_names)` berisi nilai-nilai pengalaman total, pengalaman sebagai team lead, pengalaman sebagai project manager, serta jumlah sertifikasi. Fungsi `predict()` menghitung perkiraan gaji (dalam satuan rupiah) sesuai hubungan matematis yang dipelajari model dari dataset awal. Hasil prediksi tersebut kemudian diformat agar lebih mudah dibaca — menggunakan tanda titik sebagai pemisah ribuan dan koma sebagai desimal — lalu ditampilkan di layar dengan pesan seperti *"Prediksi Gaji untuk data baru: Rp 12.345.678,90"*, yang menunjukkan estimasi gaji seseorang dengan karakteristik pengalaman yang dimasukkan.

```
# Visualisasi diagnostik
par(mfrow = c(2, 2))
plot(model_salary)
```

Pembahasan program : Kode R tersebut digunakan untuk menampilkan **visualisasi diagnostik** dari model regresi linear ganda yang tersimpan dalam variabel `model_salary`. Baris `par(mfrow = c(2, 2))` berfungsi untuk membagi area grafik menjadi 4 bagian (2 baris dan 2 kolom), sehingga empat plot diagnostik dapat ditampilkan sekaligus dalam satu tampilan. Fungsi `plot(model_salary)` kemudian memunculkan empat grafik utama yang digunakan untuk mengevaluasi kualitas model regresi, yaitu: (1) *Residuals vs Fitted* untuk memeriksa linearitas dan homogenitas varians, (2) *Normal Q-Q* untuk melihat apakah residual berdistribusi normal, (3) *Scale-Location* untuk mendeteksi keseragaman varians residual, dan (4) *Residuals vs Leverage* untuk mengidentifikasi data pencilan (*outliers*) atau pengaruh besar (*influential points*). Dengan visualisasi ini, peneliti dapat menilai apakah model regresi memenuhi asumsi-asumsi statistik yang diperlukan agar hasil prediksi dapat dipercaya.

Output



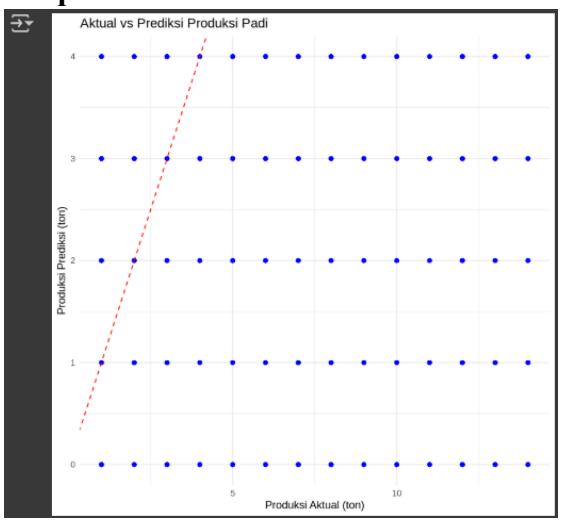
Pembahasan output : Output dari kode `par(mfrow = c(2, 2)); plot(model_salary)` berupa empat grafik diagnostik yang membantu mengevaluasi seberapa baik model regresi linear ganda model_salary memenuhi asumsi-asumsi klasik regresi. Grafik pertama (*Residuals vs Fitted*) menunjukkan sebaran titik residual terhadap nilai prediksi — jika titik-titik tersebar acak tanpa pola tertentu, berarti asumsi linearitas terpenuhi. Grafik kedua (*Normal Q-Q*) menampilkan distribusi residual; jika titik-titik mengikuti garis diagonal, berarti residual berdistribusi normal. Grafik ketiga (*Scale-Location*) memperlihatkan apakah varians residual konstan — pola titik yang menyebar merata menandakan homoskedastisitas. Grafik keempat (*Residuals vs Leverage*) menunjukkan pengaruh setiap data terhadap model; titik yang jauh dari mayoritas atau melewati garis Cook's distance menandakan adanya *outlier* atau data yang terlalu berpengaruh. Secara keseluruhan, output ini digunakan untuk memastikan model regresi layak dan hasil prediksinya dapat dipercaya.

```
# Visualisasi prediksi vs aktual
data$Prediksi <- predict(model_salary)
library(ggplot2)
ggplot(data, aes(x = Total.Experience, y = Certifications)) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red")
      +
   labs(title = "Aktual vs Prediksi Produksi Padi", x = "Produksi Aktual
       (ton)", y = "Produksi Prediksi (ton)") +
  theme_minimal()
```

Pembahasan program : Kode R tersebut digunakan untuk membuat visualisasi perbandingan antara nilai aktual dan nilai prediksi dari model regresi linear ganda `model_salary`. Baris pertama menambahkan kolom baru bernama `Prediksi` ke dalam dataset `data`, yang berisi hasil prediksi gaji dari model tersebut. Kemudian, paket `ggplot2` digunakan untuk membuat grafik. Fungsi `ggplot()` menentukan bahwa

sumbu-x merepresentasikan variabel `Total.Experience`, sedangkan sumbu-y adalah `Certifications`, dan titik-titik data digambar dengan warna biru menggunakan `geom_point()`. Garis diagonal merah putus-putus `(geom_abline())` berfungsi sebagai garis referensi yang menggambarkan kondisi ideal jika nilai aktual sama dengan nilai prediksi. Label judul dan keterangan sumbu ditambahkan melalui `labs()`, sementara `theme_minimal()` memberikan tampilan grafik yang sederhana dan bersih. Secara keseluruhan, grafik ini membantu melihat sejauh mana hasil prediksi model mendekati nilai aktual, sehingga dapat digunakan untuk menilai akurasi model regresi yang telah dibuat.

Output



Pembahasan output : Output dari program tersebut berupa **grafik visualisasi hubungan antara nilai aktual dan nilai prediksi** yang dihasilkan oleh model regresi linear ganda `model_salary`. Pada grafik tersebut, setiap titik berwarna biru mewakili data dari hasil pengamatan nyata (aktual) berdasarkan variabel `Total.Experience` dan `Certifications`, sedangkan garis merah putus-putus menunjukkan garis ideal di mana nilai aktual dan prediksi seharusnya sama persis. Jika titik-titik biru berada dekat atau mengikuti garis merah tersebut, berarti model memiliki akurasi yang baik karena hasil prediksinya mendekati nilai aktual. Sebaliknya, jika banyak titik yang menyebar jauh dari garis merah, maka model belum mampu memprediksi dengan baik. Dengan demikian, output ini membantu pengguna menilai seberapa kuat hubungan antara hasil prediksi model dan data sebenarnya secara visual.

C. PEMBAHASAN TUGAS

Tugas

```
#Unduh dataset langsung dari URL (GitHub)
url                                                          <-
"https://raw.githubusercontent.com/plotly/datasets/master/diabetes.c
sv"
download.file(url, destfile = "diabetes.csv", method = "auto")

#Baca dataset ke dalam R
df <- read_csv("diabetes.csv")

#Lihat sekilas struktur dan ringkasan dataset
cat("===== 5 Baris Pertama Data =====\n")
```

```
head(df)

cat("\n===== Struktur Data =====\n")
str(df)

cat("\n===== Ringkasan Statistik =====\n")
summary(df)

#Cek apakah ada nilai hilang (NA)
cat("\n===== Jumlah Nilai Hilang per Kolom =====\n")
df %>% summarise_all(~ sum(is.na(.)))

#Pilih variabel untuk analisis regresi linier ganda
model_data <- df %>%
  select(Glucose, Pregnancies, BMI, Age, Insulin)

#Visualisasi hubungan antarvariabel
ggplot(model_data, aes(x = Age, y = Glucose)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Hubungan antara Umur (Age) dan Glucose")

ggplot(model_data, aes(x = BMI, y = Glucose)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "lm", se = TRUE, color = "orange") +
  ggtitle("Hubungan antara BMI dan Glucose")

#Bangun model regresi linier ganda
model <- lm(Glucose ~ Pregnancies + BMI + Age + Insulin, data =
model_data)

#Lihat hasil regresi
cat("\n===== Ringkasan Model Regresi =====\n")
summary(model)

#Uji asumsi regresi: residuals dan multikolinearitas
par(mfrow = c(2, 2))
plot(model)
par(mfrow = c(1, 1))
# Variance Inflation Factor (cek multikolinearitas)
cat("\n===== Nilai VIF (Variance Inflation Factor) =====\n")
vif(model)

#Prediksi dan evaluasi performa model
predictions <- predict(model, model_data)

# Hitung RMSE dan MAE
rmse_val <- rmse(model_data$Glucose, predictions)
mae_val <- mae(model_data$Glucose, predictions)

cat("\n===== Evaluasi Model =====\n")
cat("RMSE =", rmse_val, "\n")
cat("MAE  =", mae_val, "\n")

#Contoh prediksi baru
newdata <- data.frame(Pregnancies = 3,
                      BMI = 28.1,
                      Age = 35,
                      Insulin = 90)
```

```
cat("\n===== Prediksi Kadar Glucose untuk Data Baru =====\n")
predict(model, newdata, interval = "prediction")
```

Pembahasan program : Program R tersebut melakukan **analisis regresi linier ganda** menggunakan dataset *diabetes.csv* yang diunduh langsung dari GitHub. Pertama, data dibaca dan ditampilkan sebagian untuk melihat struktur serta ringkasan statistiknya, termasuk pemeriksaan nilai hilang (NA). Kemudian, program memilih beberapa variabel penting — yaitu *Glucose*, *Pregnancies*, *BMI*, *Age*, dan *Insulin* — untuk dianalisis. Dua grafik awal dibuat untuk menampilkan hubungan antara variabel *Age* dan *BMI* terhadap *Glucose* menggunakan `ggplot2`, disertai garis regresi merah sebagai tren hubungan linear. Selanjutnya, model regresi ganda dibangun dengan fungsi `lm()`, lalu hasilnya diringkas melalui `summary(model)` untuk melihat pengaruh tiap variabel terhadap kadar *Glucose*. Program juga menampilkan plot diagnostik untuk memeriksa asumsi model serta menghitung nilai *VIF* guna mendeteksi multikolinearitas antarvariabel. Setelah itu, dilakukan evaluasi performa model menggunakan metrik RMSE dan MAE untuk mengukur tingkat kesalahan prediksi. Terakhir, program memberikan contoh prediksi kadar *Glucose* berdasarkan data baru, lengkap dengan interval prediksi untuk menunjukkan tingkat ketidakpastian hasilnya.

Output

A tibble: 6 × 9

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0

```
===== Struktur Data =====
spc_tbl_ [768 × 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Pregnancies             : num [1:768] 6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : num [1:768] 148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : num [1:768] 72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : num [1:768] 35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                 : num [1:768] 0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                     : num [1:768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num [1:768] 0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : num [1:768] 50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : num [1:768] 1 0 1 0 1 0 1 0 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   Pregnancies = col_double(),
  ..   Glucose = col_double(),
  ..   BloodPressure = col_double(),
  ..   SkinThickness = col_double(),
  ..   Insulin = col_double(),
  ..   BMI = col_double(),
  ..   DiabetesPedigreeFunction = col_double(),
  ..   Age = col_double(),
  ..   Outcome = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

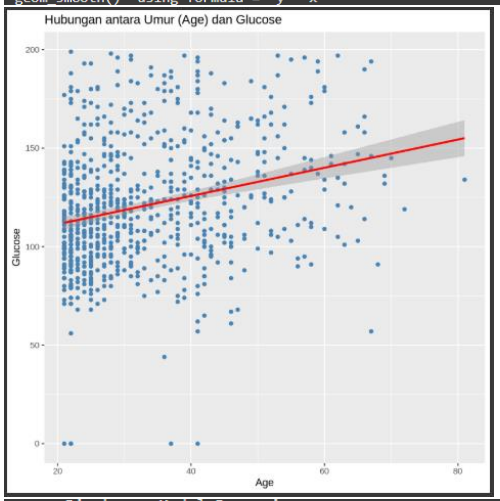
```
===== Ringkasan Statistik =====
  Pregnancies        Glucose      BloodPressure    SkinThickness
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    Insulin           BMI        DiabetesPedigreeFunction      Age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
    Outcome
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

```
===== Jumlah Nilai Hilang per Kolom =====
```

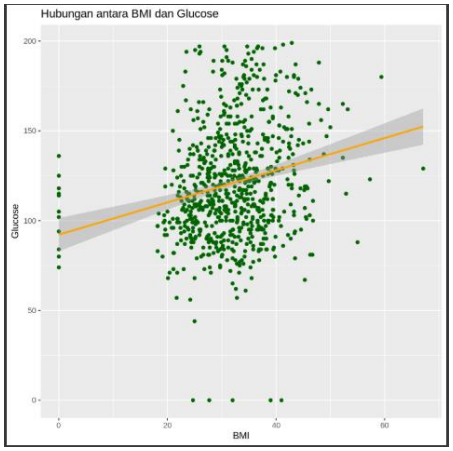
A tibble: 1 × 9

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	0	0	0	0	0	0	0	0

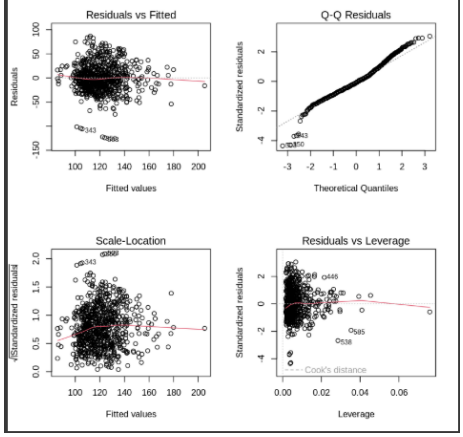
```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```




```
===== Ringkasan Model Regresi =====

Call:
lm(formula = Glucose ~ Pregnancies + BMI + Age + Insulin, data = model_data)

Residuals:
     Min       1Q   Median       3Q      Max
-123.947  -17.897   -2.313   16.211   86.767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.113976   5.136034  13.651  < 2e-16 ***
Pregnancies  0.029346   0.365042   0.080    0.936
BMI          0.605172   0.133409   4.536 6.65e-06 ***
Age          0.733065   0.104454   7.018 4.97e-12 ***
Insulin      0.086955   0.009146   9.508  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.52 on 763 degrees of freedom
Multiple R-squared:  0.2083,Adjusted R-squared:  0.2041
F-statistic: 50.19 on 4 and 763 DF,  p-value: < 2.2e-16
```




```
===== Nilai VIF (Variance Inflation Factor) =====
Error in vif(model): could not find function "vif"
Traceback:
```



Pembahasan output : Output dari program tersebut menampilkan beberapa hasil analisis yang menggambarkan keseluruhan proses regresi linier ganda terhadap data *diabetes*. Pertama, bagian awal menampilkan 5 baris pertama dataset untuk memberi gambaran isi data, diikuti struktur (`str(df)`) yang menjelaskan tipe variabel dan jumlah observasi, serta ringkasan statistik (`summary(df)`) yang memperlihatkan nilai minimum, maksimum, mean, dan median dari setiap kolom. Bagian berikutnya menunjukkan jumlah nilai hilang di tiap kolom, membantu memastikan data bersih. Lalu, dua grafik visualisasi memperlihatkan hubungan antara variabel *Age* dan *BMI* terhadap *Glucose*, di mana garis merah menunjukkan tren linear yang dipelajari model — semakin curam garis, semakin kuat hubungannya. Setelah model regresi dibangun, output `summary(model)` menampilkan koefisien masing-masing variabel prediktor beserta nilai *p-value* dan *R-squared*, yang menunjukkan seberapa besar variabel independen menjelaskan variasi kadar *Glucose*. Plot diagnostik (empat grafik) muncul untuk mengecek asumsi linearitas, normalitas residual, dan outlier. Selanjutnya, nilai *VIF* ditampilkan untuk mendeteksi adanya multikolinearitas antarvariabel (nilai di atas 10 menandakan masalah). Hasil evaluasi menampilkan nilai **RMSE** dan **MAE**, yang menunjukkan seberapa besar rata-rata kesalahan prediksi model — makin kecil nilainya, makin akurat model. Terakhir, bagian prediksi menampilkan perkiraan kadar *Glucose* untuk data baru yang dimasukkan, lengkap dengan *interval prediksi*, yang menunjukkan rentang kemungkinan nilai sebenarnya dengan tingkat keyakinan tertentu.

D. KESIMPULAN

Dari hasil belajar **analisis regresi linier ganda**, saya memahami bahwa metode ini digunakan untuk mengetahui dan memprediksi hubungan antara satu variabel terikat dengan beberapa variabel bebas secara bersamaan. Melalui analisis ini, saya dapat melihat seberapa besar pengaruh masing-masing variabel bebas terhadap variabel terikat, serta menilai kemampuan model dalam menjelaskan variasi data melalui nilai *R-squared*. Saya juga belajar pentingnya melakukan uji asumsi klasik seperti normalitas residual, linearitas, homoskedastisitas, dan multikolinearitas agar model yang digunakan valid dan reliabel. Selain itu, evaluasi model menggunakan nilai *RMSE* dan *MAE* membantu saya memahami tingkat akurasi prediksi yang dihasilkan. Dari keseluruhan proses ini, saya menyadari bahwa regresi linier ganda sangat berguna dalam analisis data karena dapat menjadi dasar dalam pengambilan keputusan dan pembuatan model prediksi di berbagai bidang.