LAPORAN FINAL PROJECT DATA ANALYST CAMP

Nama : Nazwa Nurul Wijaya
Link Data Final :
https://docs.google.com/spreadsheets/d/1jHYizpBXRnbFZWec3szalpl6h8V2vJar/edit?usp=sharing&ouid=113236234825110356290&rtpof=true&sd=true

Tugas

1. Lakukan eksplorasi awal terhadap data pendaftaran internship.

Kode

```r
# ============================
# Final Project Data Analyst
# Nazwa Nurul Wijaya

# LOAD PACKAGES
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(stringr)

# 1. LOAD DATA
data <- Form_Pendaftaran_Internship_Jawaban_xlsx_Form_Responses_1

# 2. EKSPLORASI AWAL DATA
cat("STRUKTUR DATA:\n")
str(data)

cat("\nNAMA KOLOM:\n")
colnames(data)
```

Output

```
STRUKTUR DATA:
> str(data)
spc_tbl_ [463 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Timestamp                                                                    : chr [1:463] "1/23/2022 7:50:59" "1/23/2022 8:55:30" "1/23/2022 9:05:02" "1/23/2022 9:34:07" ...
 $ Nama Lengkap                                                                 : chr [1:463] "Aditya Pratama" "Siti Nurhaliza" "Rizky Maulana" "Dwi Lestari" ...
 $ Jenis Kelamin                                                                : chr [1:463] "Perempuan" "Perempuan" "Perempuan" "Perempuan" ...
 $ Asal Perguruan Tinggi                                                        : chr [1:463] "Universitas Padjadjaran" "Universitas Singaperbangsa Karawang" "Universitas Sumatera Utara" "Politeknik Negeri Jakarta" ...
 $ Asal Provinsi                                                                : chr [1:463] "Banten" "Jawa Barat" "Sumatera Utara" "Jawa Barat" ...
 $ Departemen apa yang kamu inginkan?                                           : chr [1:463] "Administration and Finance" "Administration and Finance" "Administration and Finance" "Administration and Finance" ...
 $ Apa alasan kamu menginginkan departemen tersebut                             : chr [1:463] "Implementasi knowledge di bidang accounting" "Alasan saya menginginkan departemen Administration and Finance yaitu kar
 $ apakah kamu bersedia untuk melakuan Internship Program di Ousean Group selama 3 bulan?: chr [1:463] "Ya, bersedia" "Ya, bersedia" "Ya, bersedia" "Ya, bersedia" ...
 - attr(*, "spec")=
  .. cols(
  ..   Timestamp = col_character(),
  ..   `Nama Lengkap` = col_character(),
  ..   `Jenis Kelamin` = col_character(),
  ..   `Asal Perguruan Tinggi` = col_character(),
  ..   `Asal Provinsi` = col_character(),
  ..   `Departemen apa yang kamu inginkan?` = col_character(),
  ..   `Apa alasan kamu menginginkan departemen tersebut` = col_character(),
  ..   `Skill apa yang kamu miliki saat ini?` = col_character(),
  ..   `Skill apa yang sedang kamu pelajari saat ini?` = col_character(),
  ..   `apakah kamu bersedia untuk melakuan Internship Program di Ousean Group selama 3 bulan?` = col_character()
  .. )

> cat("\nNAMA KOLOM:\n")

NAMA KOLOM:
> colnames(data)
 [1] "Timestamp"
 [2] "Nama Lengkap"
 [3] "Jenis Kelamin"
 [4] "Asal Perguruan Tinggi"
 [5] "Asal Provinsi"
 [6] "Departemen apa yang kamu inginkan?"
 [7] "Apa alasan kamu menginginkan departemen tersebut"
 [8] "Skill apa yang kamu miliki saat ini?"
 [9] "Skill apa yang sedang kamu pelajari saat ini?"
[10] "apakah kamu bersedia untuk melakuan Internship Program di Ousean Group selama 3 bulan?"
```

2. Tangani data kosong secara logis (imputasi/eksklusi), Identifikasi dan perbaiki data yang duplikat, kosong, atau tidak konsisten.

a. Kode

```
# Mengganti nama agar mudah dipangil
colnames(data) <- c("Timestamp", "Nama_Lengkap", "Jenis_Kelamin",
                    "Asal_Perguruan_Tinggi", "Asal_Provinsi",
                    "Departemen", "Alasan_Departemen",
                    "Skill_Milik", "Skill_Pelajari", "Bersedia_Internship")

cat("PREVIEW DATA:\n")
head(data)
cat("\n")
```

Output

```
> cat("PREVIEW DATA:\n")
PREVIEW DATA:
> head(data)
# A tibble: 6 × 10
  Timestamp      Nama_Lengkap Jenis_Kelamin Asal_Perguruan_Tinggi Asal_Provinsi Departemen Alasan_Departemen
  <chr>          <chr>        <chr>         <chr>                 <chr>         <chr>      <chr>
1 1/23/2022 7:5… Aditya Prat… Perempuan     Universitas Padjadja… Banten        Administr… Implementasi kno…
2 1/23/2022 8:5… Siti Nurhal… Perempuan     Universitas Singaper… Jawa Barat    Administr… Alasan saya meng…
3 1/23/2022 9:0… Rizky Maula… Perempuan     Universitas Sumatera… Sumatera Uta… Administr… Karena menurut s…
4 1/23/2022 9:3… Dwi Lestari  Perempuan     Politeknik Negeri Ja… Jawa Barat    Administr… Karena saya memi…
5 1/23/2022 12:… Bayu Saputra Perempuan     Universitas Brawijaya Jawa Timur    Administr… Karena saya memi…
6 1/23/2022 12:… Andini Perm… Laki-laki     Universitas Brawijaya Sumatera Uta… Administr… Saya ingin menga…
# ℹ 3 more variables: Skill_Milik <chr>, Skill_Pelajari <chr>, Bersedia_Internship <chr>
> cat("\n")
```

b. Kode

```
# 3. HANDLING DATA KOSONG DAN DUPLIKAT
# Cek data kosong
cat("DATA KOSONG PER KOLOM:\n")
missing_data <- colSums(is.na(data))
print(missing_data)
cat("\n")
```

Output

```
> cat("DATA KOSONG PER KOLOM:\n")
DATA KOSONG PER KOLOM:
> missing_data <- colSums(is.na(data))
> print(missing_data)
            Timestamp          Nama_Lengkap         Jenis_Kelamin Asal_Perguruan_Tinggi
                    0                     0                     0                     0
        Asal_Provinsi            Departemen     Alasan_Departemen           Skill_Milik
                    0                     0                     0                     0
       Skill_Pelajari   Bersedia_Internship
                    0                     0
```

Hasilnya semua 0, artinya tidak ada nilai kosong **(**NA**)** di dataset. Jadi semua kolom terisi penuh.

c. Kode

```
# Cek data duplikat
cat("JUMLAH DATA DUPLIKAT:", sum(duplicated(data)), "\n\n")
```

Output

```
> # Cek data duplikat
> cat("JUMLAH DATA DUPLIKAT:", sum(duplicated(data)), "\n\n")
JUMLAH DATA DUPLIKAT: 0
```

Hasil 0 artinya tidak ada baris duplikat yang persis sama di dataset.

d. Kode

```
# Mengecek apakah ada data yang tidak konsisten
unique(data$Jenis_Kelamin)

unique(data$Asal_Provinsi)

unique(data$Departemen)
```

Output

```
> unique(data$Jenis_Kelamin)
[1] "Perempuan" "Laki-laki"
>
> unique(data$Asal_Provinsi)
 [1] "Banten"                  "Jawa Barat"              "Sumatera Utara"
 [4] "Jawa Timur"              "DKI Jakarta"             "Jawa Tengah"
 [7] "Riau"                    "Sumatera Barat"          "DI Yogyakarta"
[10] "Sumatera Selatan"        "Bali"                    "Provinsi Kalimantan Selatan"
[13] "Nusa Tenggara Timur"     "Jambi"                   "Sulawesi Selatan"
[16] "Kalimantan Tengah"       "Aceh"                    "Kalimantan Barat"
[19] "Kalimantan Timur"        "Lampung"                 "Sulawesi Utara"
[22] "Nusa Tenggara Barat"     "Bengkulu"                "Sulawesi Tenggara"
[25] "Kepulauan Bangka Belitung" "Kepulauan Riau"
```

Untuk Jenis_Kelamin Sudah konsisten hanya 2 kategori, penulisannya seragam.

Untuk Asal_Provinsi Satu saja yang kurang konsisten yaitu "Provinsi Kalimantan Selatan" sehingga saya ganti menjadi "Kalimantan Selatan" agar format sama dengan provinsi lain.

```
> unique(data$Departemen)
[1] "Administration and Finance"          "Graphic Designer"
[3] "Human Resource and Development (HRD)" "Marketing"
[5] "Project Officer"                     "Public Relation"
[7] "Social Media Officer"                "Video Editor"
[9] "WordPress Developer"
```

Semua sudah rapi dan konsisten 9 departemen, penulisan seragam.

e. Kode perbaikan yangperlu yaitu ketidak konsistenan Asal_Provinsi

```
#Membenahi ketidak konsistenan 1 nama Asal_Provinsi
data$Asal_Provinsi <- ifelse(data$Asal_Provinsi == "Provinsi Kalimantan Selatan",
                             "Kalimantan Selatan",
                             data$Asal_Provinsi)
```

Ouput

```
> data$Asal_Provinsi <- ifelse(data$Asal_Provinsi == "Provinsi Kalimantan Selatan",
+                              "Kalimantan Selatan",
+                              data$Asal_Provinsi)
> unique(data$Asal_Provinsi)
 [1] "Banten"                  "Jawa Barat"              "Sumatera Utara"
 [4] "Jawa Timur"              "DKI Jakarta"             "Jawa Tengah"
 [7] "Riau"                    "Sumatera Barat"          "DI Yogyakarta"
[10] "Sumatera Selatan"        "Bali"                    "Kalimantan Selatan"
[13] "Nusa Tenggara Timur"     "Jambi"                   "Sulawesi Selatan"
[16] "Kalimantan Tengah"       "Aceh"                    "Kalimantan Barat"
[19] "Kalimantan Timur"        "Lampung"                 "Sulawesi Utara"
[22] "Nusa Tenggara Barat"     "Bengkulu"                "Sulawesi Tenggara"
[25] "Kepulauan Bangka Belitung" "Kepulauan Riau"
```

3. Hitung statistik deskriptif:
    1. Total peserta

Kode dan Ouput

```
> # 1. Total peserta
> total_peserta <- nrow(data)
> cat("TOTAL PESERTA:", total_peserta, "\n\n")
TOTAL PESERTA: 463
```

    2. Jenis kelamin

Kode dan Output

```
80  # 2. Jenis kelamin
81  gender_stats <- data %>%
82    count(Jenis_Kelamin) %>%
83    mutate(Persentase = round(n / total_peserta * 100, 2))
84
85  cat("DISTRIBUSI JENIS KELAMIN:\n")
86  print(gender_stats)
87  cat("\n")
```

```
> # 2. Jenis kelamin
> gender_stats <- data %>%
+   count(Jenis_Kelamin) %>%
+   mutate(Persentase = round(n / total_peserta * 100, 2))
> cat("DISTRIBUSI JENIS KELAMIN:\n")
DISTRIBUSI JENIS KELAMIN:
> print(gender_stats)
# A tibble: 2 × 3
  Jenis_Kelamin     n Persentase
  <chr>         <int>      <dbl>
1 Laki-laki       125         27
2 Perempuan       338         73
```

    3. Asal provinsi

Kode dan Output

```
89  # 3. Asal provinsi
90  province_stats <- data %>%
91    count(Asal_Provinsi) %>%
92    arrange(desc(n)) %>%
93    mutate(Persentase = round(n / total_peserta * 100, 2))
94
95  cat("DISTRIBUSI ASAL PROVINSI (TOP 10):\n")
96  print(province_stats %>% head(10))
97  cat("\n")
```

```
> print(province_stats %>% head(10))
# A tibble: 10 × 3
   Asal_Provinsi        n Persentase
   <chr>            <int>      <dbl>
 1 Jawa Barat         101       21.8
 2 Jawa Timur          91       19.6
 3 Jawa Tengah         68       14.7
 4 DKI Jakarta         56       12.1
 5 Banten              55       11.9
 6 Sumatera Utara      23        4.97
 7 DI Yogyakarta       12        2.59
 8 Sumatera Barat       9        1.94
 9 Sulawesi Selatan     7        1.51
10 Sumatera Selatan     5        1.08
> cat("\n")
```

4. Perguruan tinggi

Kode dan Output

```
98
99  # 4. Perguruan tinggi
100 university_stats <- data%>%
101   count(Asal_Perguruan_Tinggi) %>%
102   arrange(desc(n)) %>%
103   mutate(Persentase = round(n / total_peserta * 100, 2))
104
105 cat("DISTRIBUSI PERGURUAN TINGGI (TOP 10):\n")
106 print(university_stats %>% head(10))
107 cat("\n")
```

```
# A tibble: 10 × 3
   Asal_Perguruan_Tinggi                n Persentase
   <chr>                            <int>      <dbl>
 1 Universitas Diponegoro              29       6.26
 2 Universitas Brawijaya               27       5.83
 3 Universitas Airlangga               18       3.89
 4 Universitas Sebelas Maret           11       2.38
 5 Universitas Indonesia                8       1.73
 6 UIN Syarif Hidayatullah Jakarta      7       1.51
 7 Universitas Negeri Malang            7       1.51
 8 Universitas Negeri Yogyakarta        7       1.51
 9 Universitas Padjadjaran              7       1.51
10 Universitas Pamulang                 7       1.51
> cat("\n")
```

5. Departemen yang diminati

Kode dan Output

```
109 # 5. Departemen yang diminati
110 department_stats <- data %>%
111   count(Departemen) %>%
112   arrange(desc(n)) %>%
113   mutate(Persentase = round(n / total_peserta * 100, 2))
114
115 cat("DISTRIBUSI DEPARTEMEN:\n")
116 print(department_stats)
117 cat("\n")
```

```
DISTRIBUSI DEPARTEMEN:
> print(department_stats)
# A tibble: 9 × 3
  Departemen                          n Persentase
  <chr>                           <int>      <dbl>
1 Administration and Finance        123       26.6
2 Human Resource and Development (HRD)  119   25.7
3 Public Relation                    58       12.5
4 Social Media Officer               47       10.2
5 Project Officer                    43        9.29
6 Marketing                          41        8.86
7 WordPress Developer                17        3.67
8 Graphic Designer                   14        3.02
9 Video Editor                        1        0.22
> cat("\n")
```

6. Hitung metrik unik: Rasio Pria vs Wanita per Departemen dan Distribusi Peserta per Kota serta Skill

a. Kode

```
# Rasio Pria vs Wanita per Departemen
gender_ratio_dept <- data %>%
  group_by(Departemen, Jenis_Kelamin) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = Jenis_Kelamin, values_from = Count, values_fill = 0) %>%
  mutate(
    Total = `Laki-laki` + Perempuan,
    Rasio_Pria = round(`Laki-laki` / Total * 100, 2),
    Rasio_Wanita = round(Perempuan / Total * 100, 2)
  )

cat("RASIO GENDER PER DEPARTEMEN:\n")
print(gender_ratio_dept)
cat("\n")
```

Output

```
RASIO GENDER PER DEPARTEMEN:
> print(gender_ratio_dept)
# A tibble: 9 × 6
  Departemen                          `Laki-laki` Perempuan Total Rasio_Pria Rasio_Wanita
  <chr>                                     <int>     <int> <int>       <dbl>        <dbl>
1 Administration and Finance                   22       101   123        17.9         82.1
2 Graphic Designer                              2        12    14        14.3         85.7
3 Human Resource and Development (HRD)         32        87   119        26.9         73.1
4 Marketing                                    20        21    41        48.8         51.2
5 Project Officer                              13        30    43        30.2         69.8
6 Public Relation                              17        41    58        29.3         70.7
7 Social Media Officer                          9        38    47        19.2         80.8
8 Video Editor                                  0         1     1           0          100
9 WordPress Developer                          10         7    17        58.8         41.2
> cat("\n")
```

b. Kode

```
# Analisis Skill (mengambil dari kolom Skill_Milik)
# Ekstrak semua skill dari kolom Skill_Milik
all_skills <- data %>%
  select(Skill_Milik) %>%
  mutate(Skill_Milik = str_split(tolower(Skill_Milik), ",")) %>%
  unnest(Skill_Milik) %>%
  mutate(Skill_Milik = str_trim(Skill_Milik)) %>%
  filter(Skill_Milik != "") %>%
  count(Skill_Milik) %>%
  arrange(desc(n)) %>%
  mutate(Persentase = round(n / total_peserta * 100, 2))

cat("DISTRIBUSI SKILL (TOP 15):\n")
print(all_skills %>% head(15))
cat("\n")
```

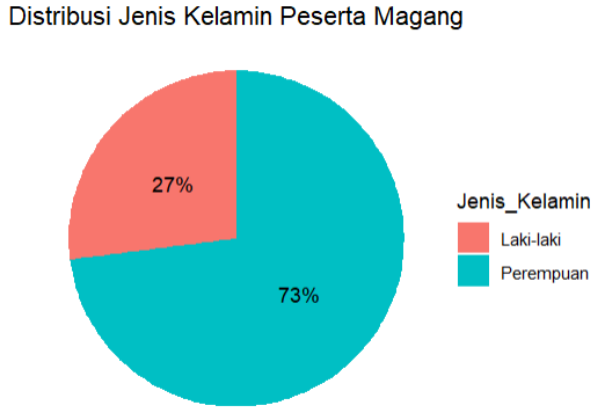
Output

```
> cat("DISTRIBUSI SKILL (TOP 15):\n")
DISTRIBUSI SKILL (TOP 15):
> print(all_skills %>% head(15))
# A tibble: 15 × 3
   Skill_Milik            n Persentase
   <chr>              <int>      <dbl>
 1 leadership           303       65.4
 2 public speaking      259       55.9
 3 copy writing         128       27.6
 4 editing video         77       16.6
 5 desaign graphic       72       15.6
 6 microsoft office      18        3.89
 7 communication         17        3.67
 8 teamwork              17        3.67
 9 problem solving       10        2.16
10 time management       10        2.16
11 accounting             9        1.94
12 critical thinking      8        1.73
13 komunikasi             8        1.73
14 project management     8        1.73
15 team work              6        1.3
> cat("\n")
```

VISUALISASI

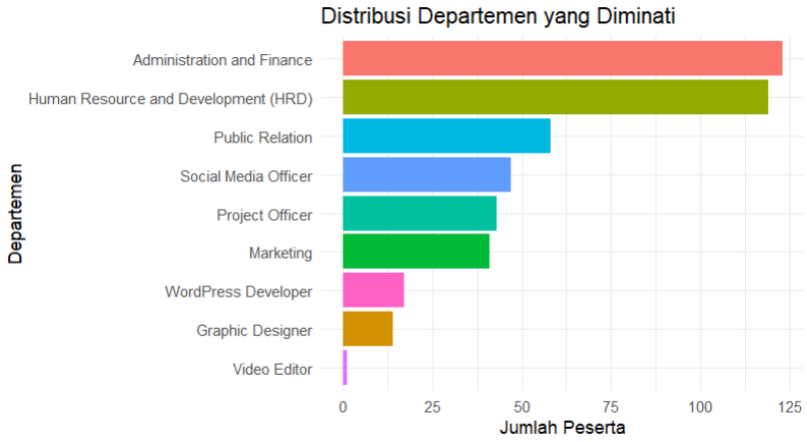
1. Kode

```r
# 6. VISUALISASI DATA
# Visualisasi Jenis Kelamin
gender_plot <- ggplot(gender_stats, aes(x = "", y = n, fill = Jenis_Kelamin)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(Persentase), "%")),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Distribusi Jenis Kelamin Peserta Magang") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

print(gender_plot)
```



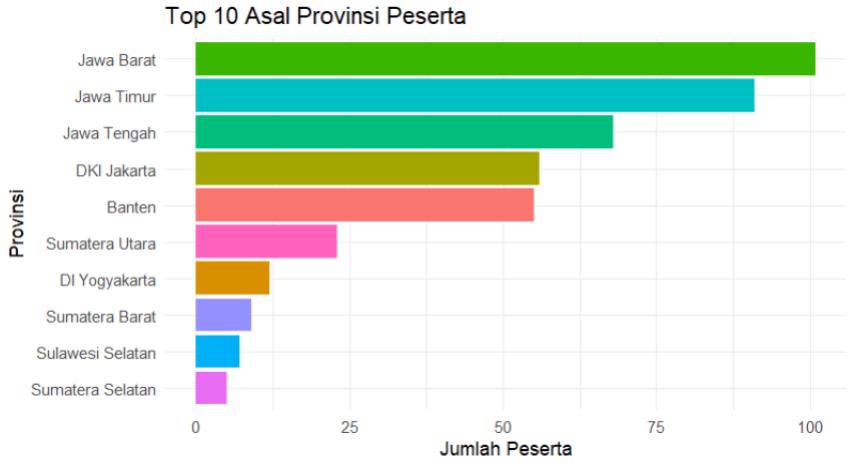
2. Kode

```r
# Visualisasi Departemen
department_plot <- ggplot(department_stats, aes(x = reorder(Departemen, n), y = n, fill = Departemen)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Distribusi Departemen yang Diminati",
       x = "Departemen", y = "Jumlah Peserta") +
  theme_minimal() +
  theme(legend.position = "none")
```



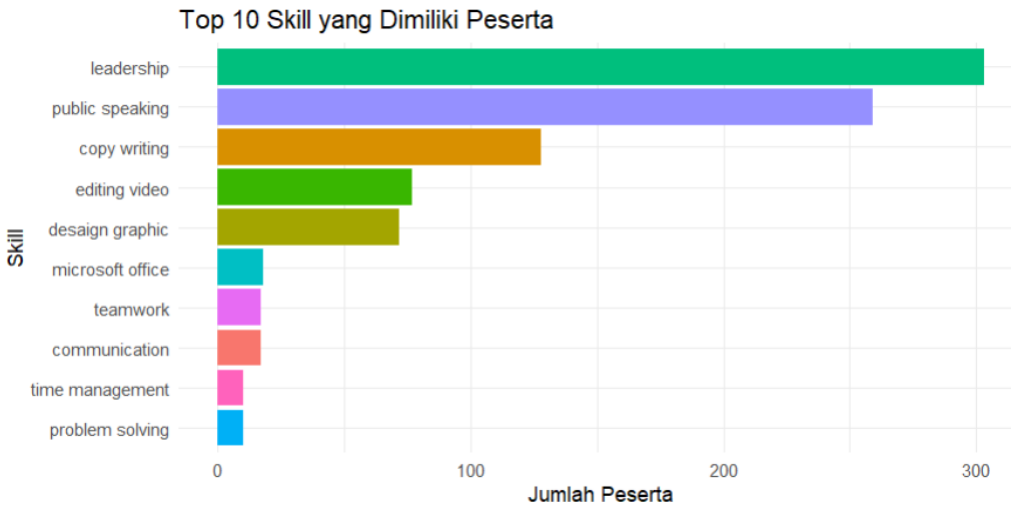
3. Kode

```r
# Visualisasi Asal Provinsi (Top 10)
province_plot <- ggplot(province_stats %>% head(10), aes(x = reorder(Asal_Provinsi, n), y = n, fill = Asal_Provinsi)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Asal Provinsi Peserta",
       x = "Provinsi", y = "Jumlah Peserta") +
  theme_minimal() +
  theme(legend.position = "none")

print(province_plot)
```



4. Kode

```r
# Visualisasi Skill (Top 10)
skills_plot <- ggplot(all_skills %>% head(10), aes(x = reorder(Skill_Milik, n), y = n, fill = Skill_Milik)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Skill yang Dimiliki Peserta",
       x = "Skill", y = "Jumlah Peserta") +
  theme_minimal() +
  theme(legend.position = "none")

print(skills_plot)
```



5. Kode

```r
# Visualisasi Rasio Gender per Departemen
gender_dept_plot <- ggplot(gender_ratio_dept, aes(x = Departemen)) +
  geom_bar(aes(y = `Laki-laki`), stat = "identity", position = "stack", fill = "blue", alpha = 0.7) +
  geom_bar(aes(y = -Perempuan), stat = "identity", position = "stack", fill = "pink", alpha = 0.7) +
  coord_flip() +
  labs(title = "Rasio Gender per Departemen",
       y = "Jumlah Peserta") +
  theme_minimal()

print(gender_dept_plot)
```