PERTEMUAN 4
JENIS GRAFIK BERDASARKAN TIPE DATA

SUB-CAPAIAN PEMBELAJARAN MATA KULIAH:

Dari proses pembelajaran pada modul 4, mahasiswa diharapkan dapat memahami mengenai jenis-jenis grafik berdasarkan tipe datanya dan membuat grafik sesuai tipe data.

POKOK BAHASAN:

1. Visualisasi dasar
2. Visualisasi atribut numerik
3. Visualisasi atribut kategorik
4. Instal dan import package ggplot
5. Visualisasi menggunakan package ggplot

TUJUAN:

1. Dapat menjelaskan jenis-jenis grafik sesuai jenis atribut/variabel
2. Dapat membuat grafik sederhana sesuai dengan jenis atribut/variabel

DASAR TEORI:

Di bagian ini, akan dibahas beberapa tipe visualisasi atau berbagai jenis plot yang dapat dibuat berdasarkan tipe data yang anda miliki. Bab 4 lebih fokus pada penyajian grafik yang efektif untuk penyampaian pesan kepada pengguna, termasuk contoh kode-kode R untuk visualisasi data sederhana menggunakan paket *ggplot*. Sebelum anda memvisualisasikan data anda, yang pertama kali harus anda lakukan adalah melakukan *import* data. R dapat melakukan proses import data dari hampir semua sumber, termasuk file teks, spreadsheet excel, paket statistik, dan sistem manajemen basis data.

a. Import data dari csv atau text file

```
library(readr)

# import data from a comma delimited file
Salaries <- read_csv("salaries.csv")

# import data from a tab delimited file
Salaries <- read_tsv("salaries.txt")
```

```
"rank","discipline","yrs.since.phd","yrs.service","sex","salary"
"Prof","B",19,18,"Male",139750
"Prof","B",20,16,"Male",173200
"AsstProf","B",4,3,"Male",79750
"Prof","B",45,39,"Male",115000
"Prof","B",40,41,"Male",141500
"AssocProf","B",6,6,"Male",97000
```

b. Import data dari excel workbooks

```
library(readxl)

# import data from an Excel workbook
Salaries <- read_excel("salaries.xlsx", sheet=1)
```

PENGANTAR GGPLOT

Bagian ini memberikan gambaran singkat tentang cara kerja paket ggplot2. Fungsi-fungsi dalam paket *ggplot* dapat digunakan untuk membuat grafik pada suatu layer. Dalam modul ini, kita akan membuat grafik sederhana menggunakan paket ggplot2 pada R. Ggplot2 memungkinkan anda untuk membuat grafik yang dapat merepresentasikan data numerik dan kategorik baik univariat maupun multivariat secara simultan. Data yang digunakan adalah data *Current Population Survey* tahun 1985 atau disingkat CPS85 yang bertujuan untuk mengeksplorasi hubungan antara upah (wage) dan pengalaman (expr). Data tersebut berisi informasi tentang upah dan karakteristik lain dari seorang pekerja, termasuk jenis kelamin, jumlah tahun pendidikan, tahun pengalaman kerja, status pekerjaan, wilayah tempat tinggal, dan keanggotaan dalam serikat pekerja. Data dapat diunduh dari halaman berikut ini:

a. Load data

```
# load data
data(CPS85 , package = "mosaicData")
```

b. Menentukan dataset dan mapping data

```
# specify dataset and mapping
library(ggplot2)
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage))
```

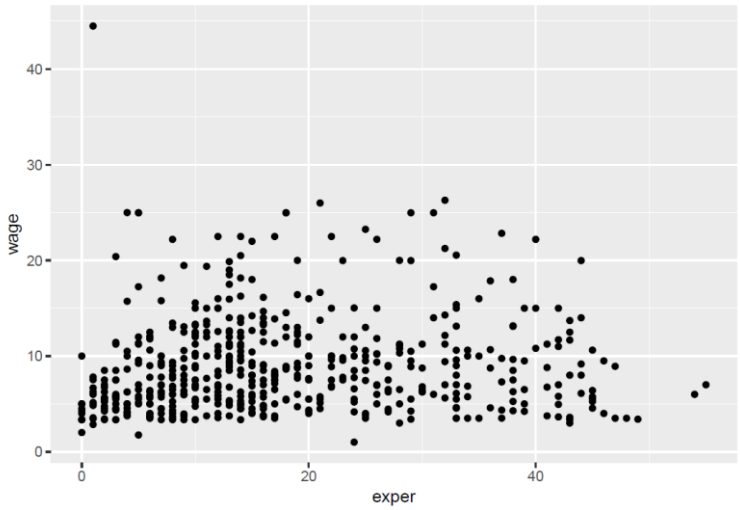
Pada tahapan ini grafik akan terlihat kosong, karena kita belum menambahkan geoms (*geometry object*).

c. Geoms (*geometry object*)

Geom adalah objek geometris (*points, lines, bars*, dll) yang dapat ditampilkan pada suatu grafik.

```
# add points
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage)) +
  geom_point()
```

Output:



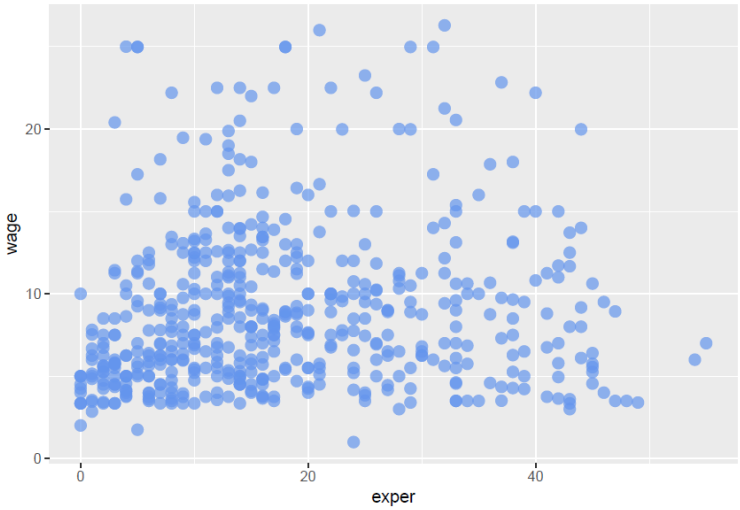
Dari grafik tersebut dapat terlihat adanya data outlier. Untuk menghapusnya dapat menggunakan perintah di bawah ini:

```
# delete outlier
library(dplyr)
plotdata <- filter(CPS85, wage < 40)
```

d. Memodifikasi *color, transparency, dan size*

```
# make points blue, larger, and semi-transparent
ggplot(data = plotdata,
       mapping = aes(x = exper, y = wage)) +
  geom_point(color = "cornflowerblue",
             alpha = .7,
             size = 3)
```

Output:

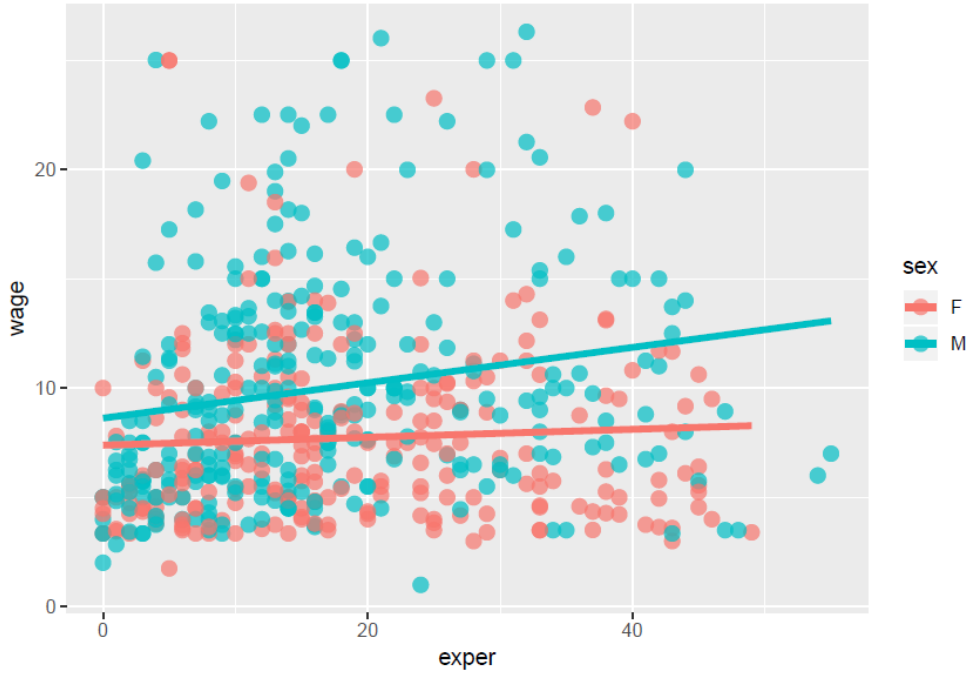


e. *Grouping*

Plotting variabel jenis kelamin dan menggambarkan variabel tersebut dengan warna.

```
# indicate sex using color
ggplot(data = plotdata,
       mapping = aes(x = exper,
                     y = wage,
                     color = sex)) +
  geom_point(alpha = .7,
             size = 3) +
  geom_smooth(method = "lm",
              se = FALSE,
              size = 1.5)
```

Output:



f. *Facets*

Facets digunakan untuk membuat grafik untuk setiap tingkat variabel tertentu (atau kombinasi variabel).

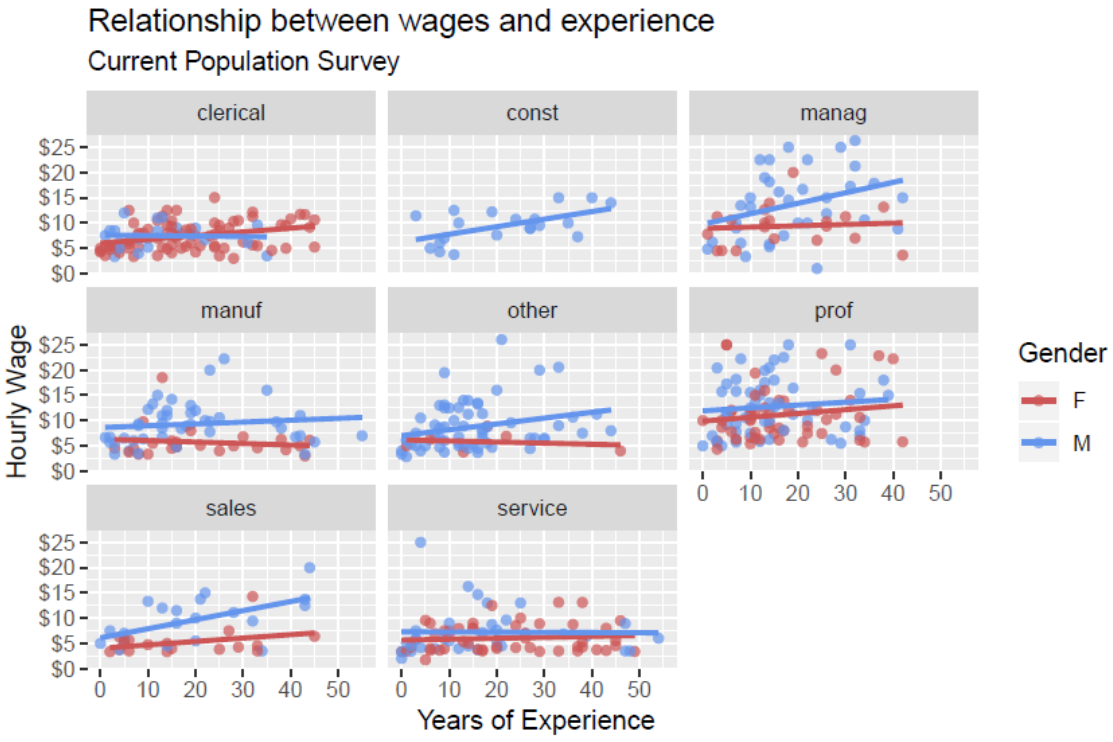
```
# reproduce plot for each level of job sector
ggplot(data = plotdata,
       mapping = aes(x = exper,
                     y = wage,
                     color = sex)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm",
              se = FALSE) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3",
                                "cornflowerblue")) +
  facet_wrap(~sector)
```

g. *Labels*

Dalam memudahkan proses interpretasi, setiap grafik harus diberikan label yang informatif.

```
# add informative labels
ggplot(data = plotdata,
       mapping = aes(x = exper,
                     y = wage,
                     color = sex)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm",
              se = FALSE) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3",
                                "cornflowerblue")) +
  facet_wrap(~sector) +
  labs(title = "Relationship between wages and experience",
       subtitle = "Current Population Survey",
       caption = "source: http://mosaic-web.org/",
       x = " Years of Experience",
       y = "Hourly Wage",
       color = "Gender")
```

Output:



JENIS-JENIS GRAFIK

Grafik Univariat

Grafik univariat melakukan *plotting* distribusi data dari hanya satu variabel. Variabel dapat bersifat kategorik (misalnya, ras, jenis kelamin) atau kuantitatif (misalnya, usia, berat badan).

1. Kategorik

   Distribusi variabel dengan kategori tunggal biasanya diplot menggunakan diagram batang dan diagram lingkaran.

   a. Bar Charts (Diagram Batang)

      Diagram batang sangat baik untuk membandingkan frekuensi kelompok yang berbeda. Seperti contoh pada dataset pernikahan berisi catatan pernikahan dari 98 orang di Mobile County, Alabama. Di bawah ini, diagram batang digunakan untuk menampilkan distribusi peserta pernikahan menurut ras.

```
library(ggplot2)
data(Marriage, package = "mosaicData")

# plot the distribution of race
ggplot(Marriage, aes(x = race)) +
  geom_bar()
```

```
# plot the distribution of race with modified colors and labels
ggplot(Marriage, aes(x = race)) +
  geom_bar(fill = "cornflowerblue",
           color="black") +
  labs(x = "Race",
       y = "Frequency",
       title = "Participants by race")
```

      Output:

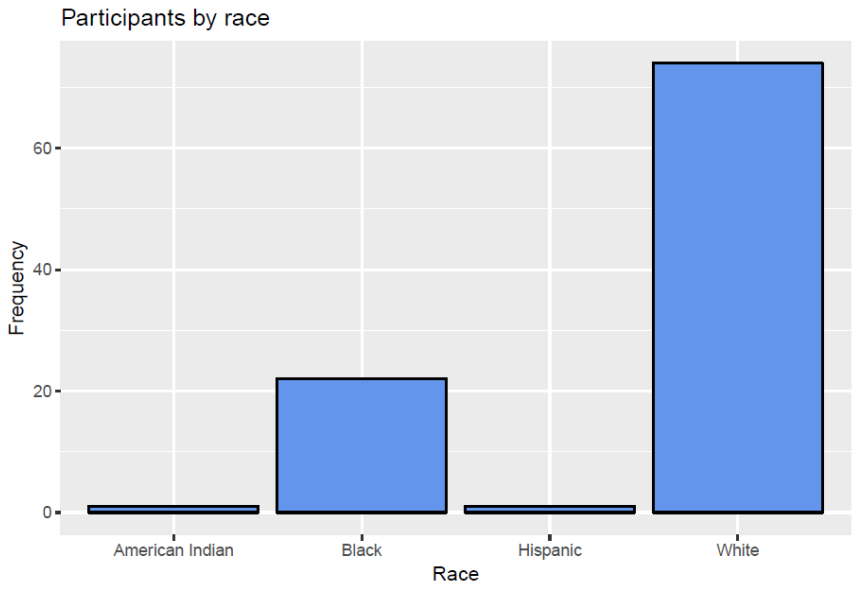
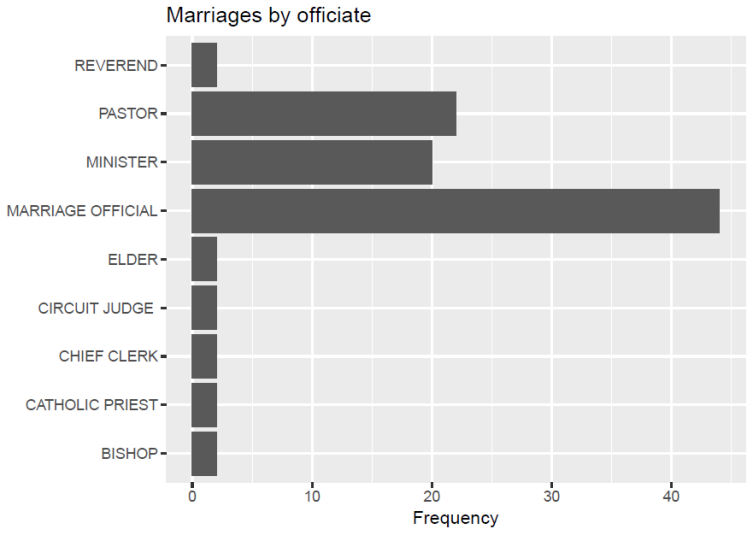


Diagram batang horizontal:

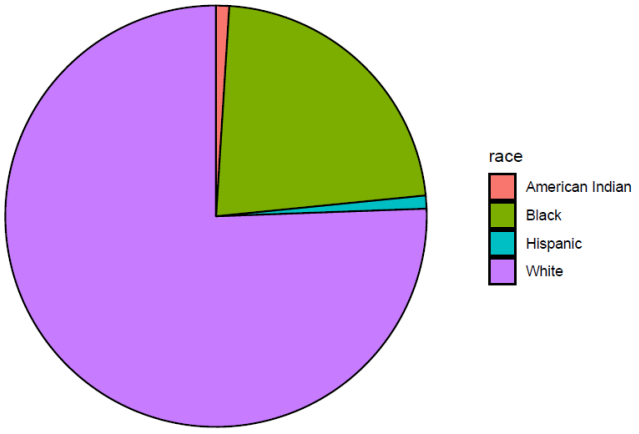
```
# horizontal bar chart
ggplot(Marriage, aes(x = officialTitle)) +
  geom_bar() +
  labs(x = "",
       y = "Frequency",
       title = "Marriages by officiate") +
  coord_flip()
```

Output:



b. Pie Chart (Diagram Lingkaran)

Diagram lingkaran kontroversial dalam statistik. Jika tujuan anda adalah untuk membandingkan frekuensi kategori, anda lebih baik menggunakan diagram batang (manusia lebih baik dalam menilai panjang batang daripada volume irisan pie). Jika tujuan anda adalah membandingkan setiap kategori secara keseluruhan dan jumlah kategorinya kecil, maka diagram lingkaran mungkin lebih cocok untuk digunakan. Untuk membuat diagram lingkaran yang menarik dengan R dibutuhkan lebih banyak penulisan kode.



```r
# create a basic ggplot2 pie chart
 plotdata <- Marriage %>%
    count(race) %>%
    arrange(desc(race)) %>%
    mutate(prop = round(n * 100 / sum(n), 1),
           lab.ypos = cumsum(prop) - 0.5  *prop)

  ggplot(plotdata,
         aes(x = "",
             y = prop,
             fill = race)) +
    geom_bar(width = 1,
             stat = "identity",
             color = "black") +
    coord_polar("y",
                start = 0,
                direction = -1) +
    theme_void()
```

2. Kuantitatif

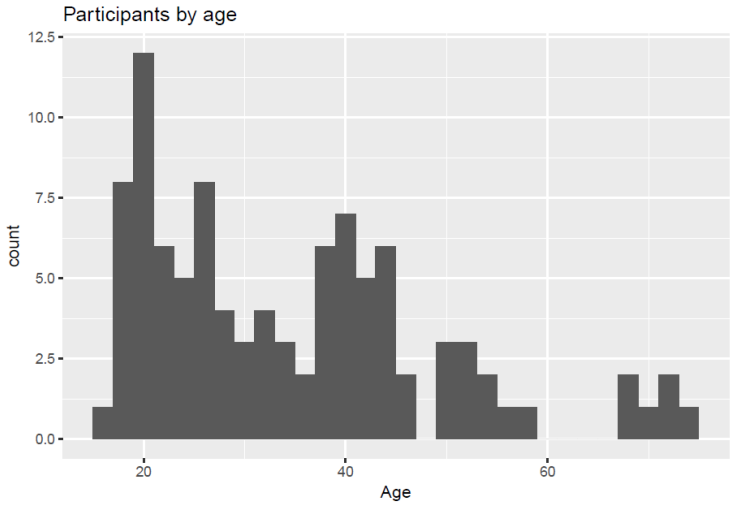
   Distribusi variabel kuantitatif tunggal biasanya diplot menggunakan histogram dan *kernel density plot* atau plot titik.

   a. Histogram

      Menggunakan dataset pernikahan, mari kita plot usia peserta pernikahan.

```r
library(ggplot2)

# plot the age distribution using a histogram
ggplot(Marriage, aes(x = age)) +
  geom_histogram() +
  labs(title = "Participants by age",
       x = "Age")
```

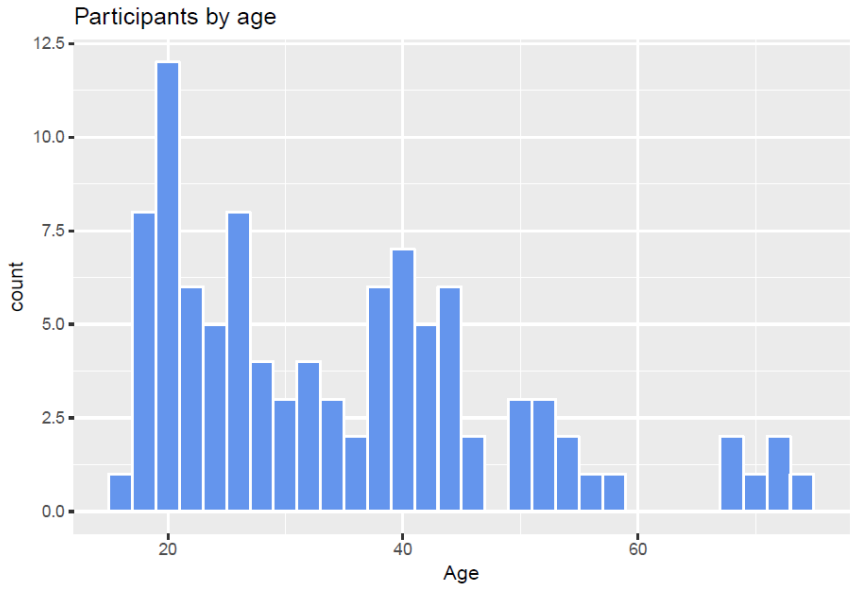
      Output:



      Sebagian besar peserta tampaknya berusia awal 20-an dengan kelompok lain berusia 40-an, dan kelompok yang jauh lebih kecil di akhir usia enam puluhan dan awal tujuh puluhan. Warna histogram dapat dimodifikasi menggunakan dua opsi: *fill* & *color*.

```
# plot the histogram with blue bars and white borders
ggplot(Marriage, aes(x = age)) +
  geom_histogram(fill = "cornflowerblue",
                 color = "white") +
  labs(title="Participants by age",
       x = "Age")
```

Output:

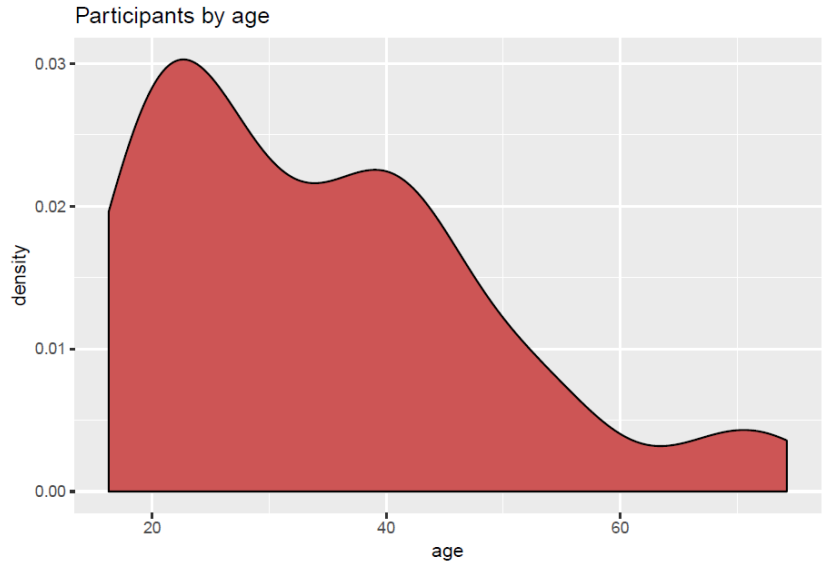


b. Kernel Density Plot

Alternatif lain dari histogram adalah *Kernel Density Plot*.

```
# Create a kernel density plot of age
ggplot(Marriage, aes(x = age)) +
  geom_density(fill = "indianred3") +
  labs(title = "Participants by age")
```

Output:



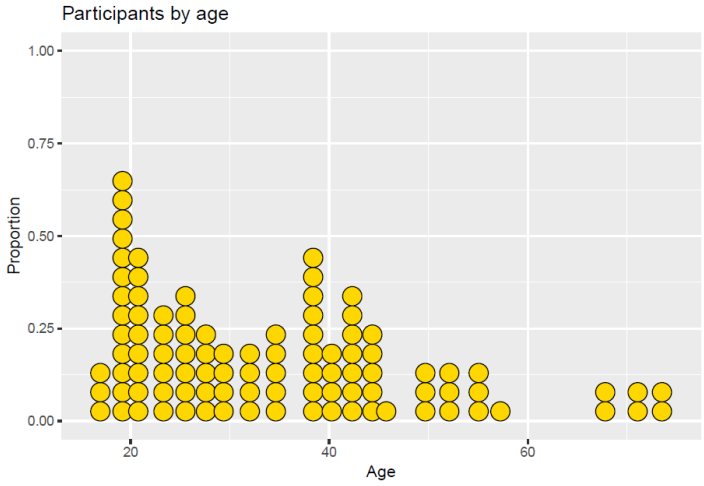
Grafik menunjukkan distribusi skor. Misalnya, proporsi kasus antara 20 dan 40 tahun akan diwakili oleh luas di bawah kurva antara 20 dan 40 pada sumbu x.

3. Dot Chart (Diagram Titik)

   Alternatif lain untuk histogram adalah diagram titik. Variabel kuantitatif dibagi ke dalam beberapa bin, berbeda dengan diagram batang, setiap pengamatan diwakili oleh sebuah titik. Diagram titik bekerja paling baik pada data dengan jumlah kecil (katakanlah, kurang dari 150).

```
# Plot ages as a dot plot using
# gold dots with black borders
ggplot(Marriage, aes(x = age)) +
  geom_dotplot(fill = "gold",
               color = "black") +
  labs(title = "Participants by age",
       y = "Proportion",
       x = "Age")
```

   Output:



Grafik Bivariat

Grafik bivariat digunakan untuk menampilkan hubungan antara dua variabel. Jenis grafik akan tergantung pada skala pengukuran variabel (kategorik atau kuantitatif).

1. *Categorical vs. Categorical*

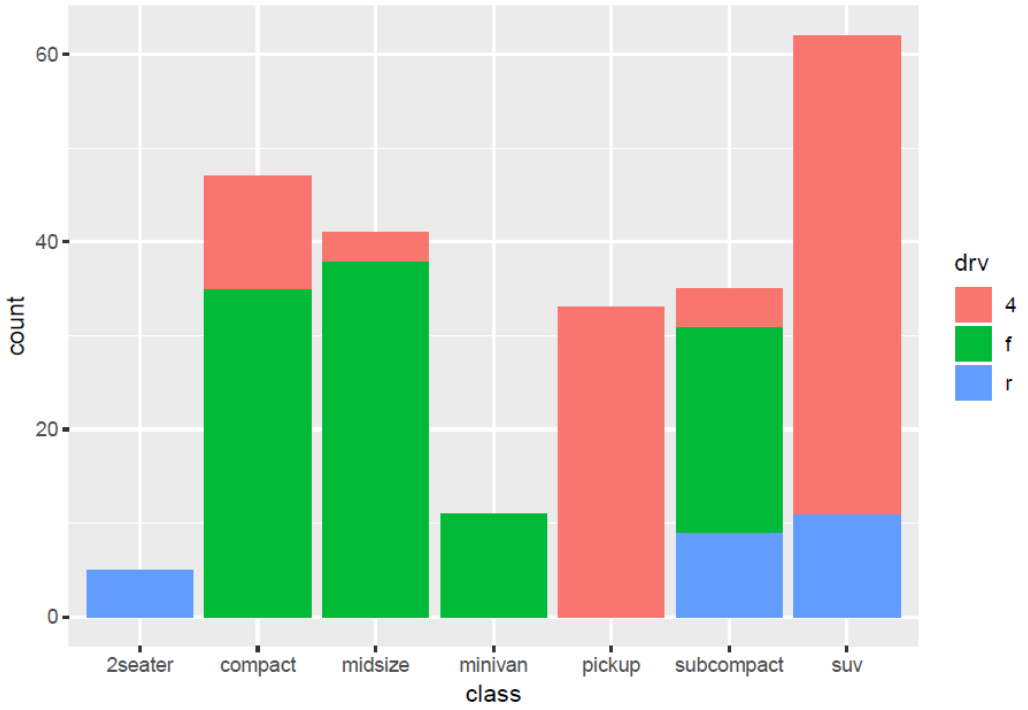
   Saat memplot hubungan antara dua variabel kategori, *stacked, grouped,* atau *segmented* bar charts dapat digunakan. Pada kali ini, kita akan menggunakan *Fuel Economy Dataset*.

   a. Stacked bar chart

      Mari kita gambarkan hubungan antara kelas mobil dan tipe penggerak (*front-wheel, rear-wheel,* atau *4-wheel drive*).

```
library(ggplot2)

# stacked bar chart
ggplot(mpg,
       aes(x = class,
           fill = drv)) +
  geom_bar(position = "stack")
```

Output:

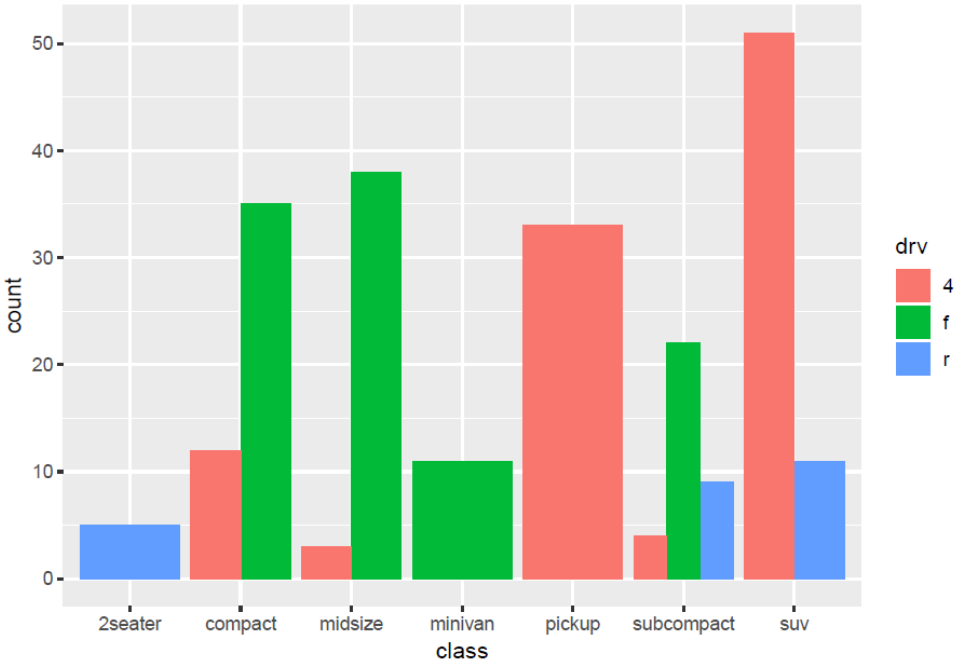


Dari grafik tersebut, kita dapat melihat misalnya, bahwa kendaraan yang paling umum adalah SUV. Semua mobil 2 kursi adalah *rear wheel drive*, sementara sebagian besar, tetapi tidak semua SUV adalah *4-wheel drive*.

b. Grouped bar chart

```
library(ggplot2)

# grouped bar plot
ggplot(mpg,
       aes(x = class,
           fill = drv)) +
  geom_bar(position = "dodge")
```

Output:

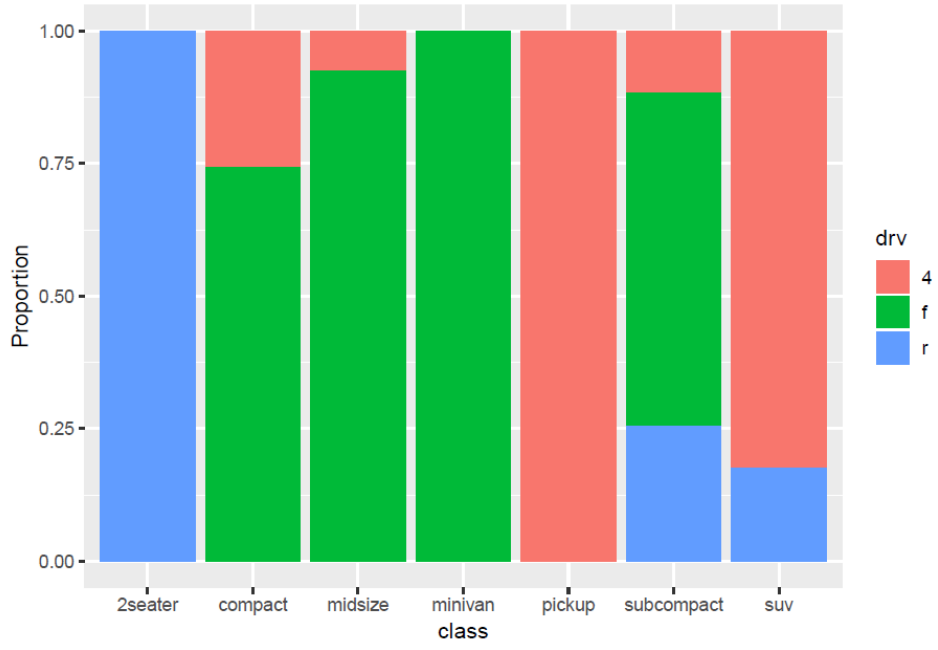


c. Segmented bar chart

Plot batang tersegmentasi adalah plot batang bertumpuk di mana setiap batang mewakili 100 persen.

```
library(ggplot2)

# bar plot, with each bar representing 100%
ggplot(mpg,
       aes(x = class,
           fill = drv)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")
```

Output:



2. *Quantitative vs. Quantitative*

Hubungan antara dua variabel kuantitatif biasanya ditampilkan menggunakan *scatterplots* dan *line graphs* (grafik garis).

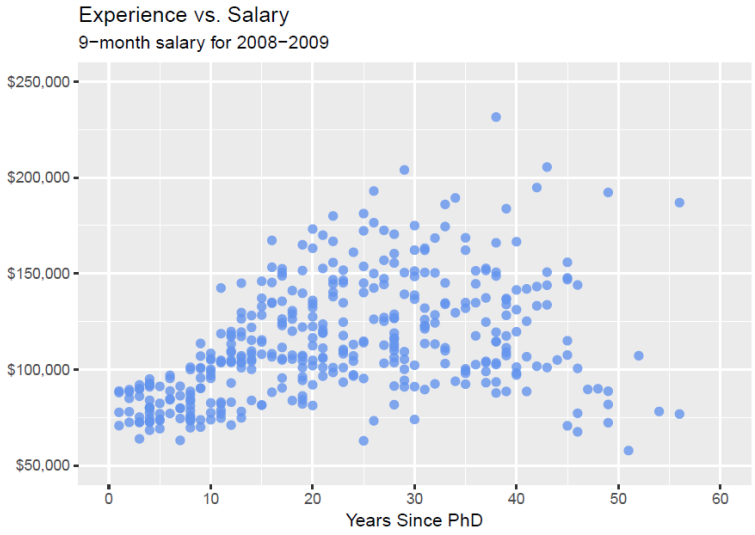
a. Scatterplot

Visualisasi paling sederhana dari dua variabel kuantitatif adalah *scatterplot*, dengan masing-masing variabel diwakili oleh sumbu x/y.

```
library(ggplot2)
data(Salaries, package="carData")
```

```
# enhanced scatter plot
ggplot(Salaries,
       aes(x = yrs.since.phd,
           y = salary)) +
  geom_point(color="cornflowerblue",
```

Output:



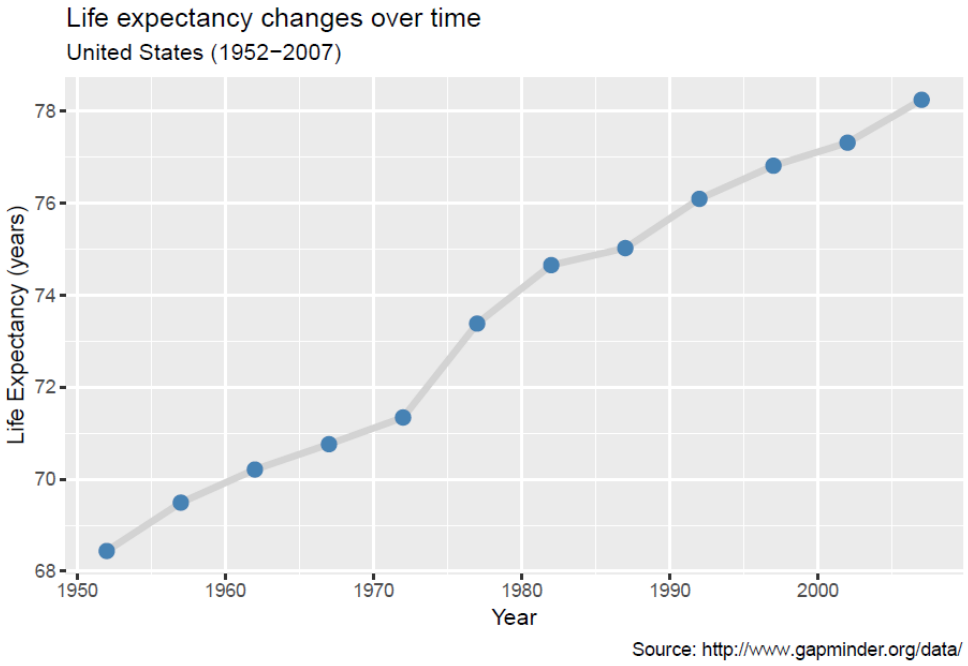
b. Line graph

Ketika salah satu dari dua variabel mewakili waktu, plot garis dapat menjadi metode yang efektif untuk menampilkan hubungan tersebut. Misalnya, kode di bawah ini menampilkan hubungan antara waktu (tahun) dan harapan hidup (lifeExp) di Amerika Serikat antara tahun 1952 dan 2007. Dataset yang digunakan adalah Gapminder.

```
data(gapminder, package="gapminder")

# Select US cases
library(dplyr)
plotdata <- filter(gapminder,
                   country == "United States")
```

```
# line plot with points
# and improved labeling
ggplot(plotdata,
       aes(x = year,
           y = lifeExp)) +
  geom_line(size = 1.5,
            color = "lightgrey") +
  geom_point(size = 3,
             color = "steelblue") +
  labs(y = "Life Expectancy (years)",
       x = "Year",
       title = "Life expectancy changes over time",
       subtitle = "United States (1952-2007)",
       caption = "Source: http://www.gapminder.org/data/")
```

Output:



3. *Categorical vs. Quantitative*

Banyak teknik visualisasi yang bisa digunakan untuk kasus data kategorik dan kuantitatif, misalnya *bar chart* dengan *summary statistics, grouped kernel density plots, side-by-side box plots, side-by-side violin plots, mean/sem plots, ridgeline plots,* dan *Cleveland plots*. Pada modul ini, akan dijelaskan mengenai *summary statistics, grouped kernel density plots, side-by-side box plots,* dan *side-by-side violin plots* saja.

a. Bar chart (*on summary statistics*)

Pada bagian sebelumnya, diagram batang digunakan untuk menampilkan jumlah kasus berdasarkan kategori untuk satu variabel maupun untuk dua variabel. Anda juga dapat menggunakan diagram batang untuk menampilkan ringkasan statistik (mis., untuk menampilkan nilai *mean* atau *median*) pada variabel kuantitatif untuk setiap kategori.

Sebagai contoh, grafik berikut menampilkan gaji rata-rata seorang profesor di universitas berdasarkan peringkat akademiknya.

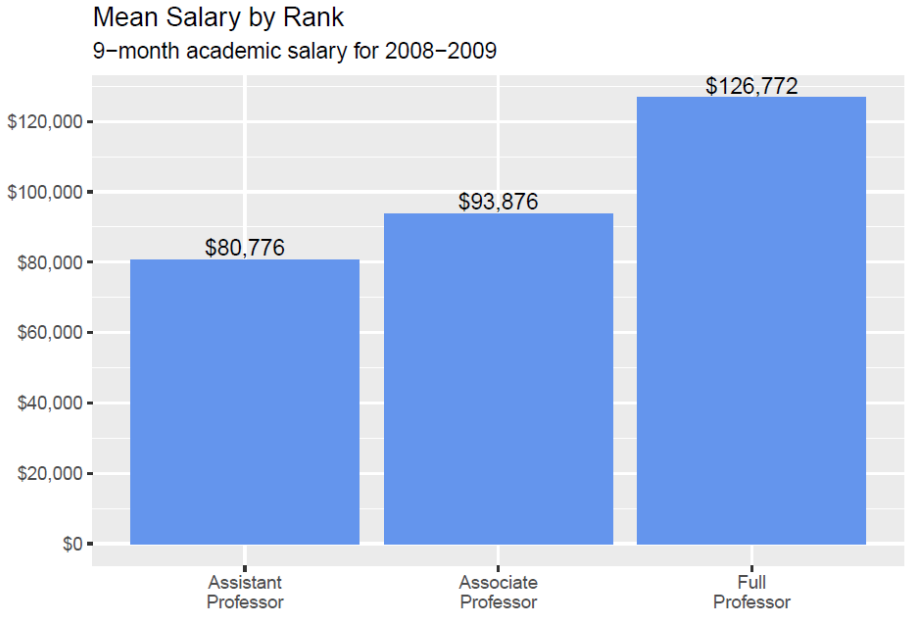
```
data(Salaries, package="carData")

# calculate mean salary for each rank
library(dplyr)
plotdata <- Salaries %>%
  group_by(rank) %>%
  summarize(mean_salary = mean(salary))
```

```
# plot mean salaries
ggplot(plotdata,
       aes(x = rank,
           y = mean_salary)) +
  geom_bar(stat = "identity")
```

```
# plot mean salaries in a more attractive fashion
library(scales)
ggplot(plotdata,
       aes(x = factor(rank,
                      labels = c("Assistant\nProfessor",
                                 "Associate\nProfessor",
                                 "Full\nProfessor")),
                      y = mean_salary)) +
  geom_bar(stat = "identity",
           fill = "cornflowerblue") +
  geom_text(aes(label = dollar(mean_salary)),
            vjust = -0.25) +
  scale_y_continuous(breaks = seq(0, 130000, 20000),
                     label = dollar) +
```

```
 labs(title = "Mean Salary by Rank",
      subtitle = "9-month academic salary for 2008-2009",
      x = "",
      y = "")
```

Output:



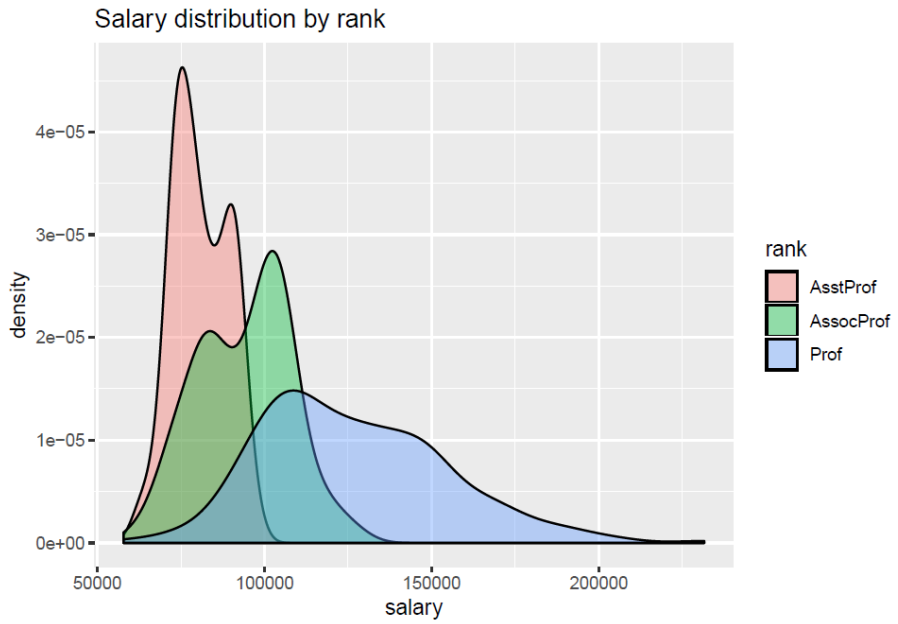
Terdapat batasan untuk jenis visualisasi ini misalnya tidak dapat menampilkan distribusi data, melainkan hanya ringkasan statistiknya saja untuk setiap kelompok.

b. Grouped kernel density plots

Anda dapat menggunakan *grouped kernel density plot* untuk membandingkan beberapa grup dengan variabel numerik dalam satu grafik.

```
# plot the distribution of salaries
# by rank using kernel density plots
ggplot(Salaries,
       aes(x = salary,
           fill = rank)) +
  geom_density(alpha = 0.4) +
  labs(title = "Salary distribution by rank")
```

Output:

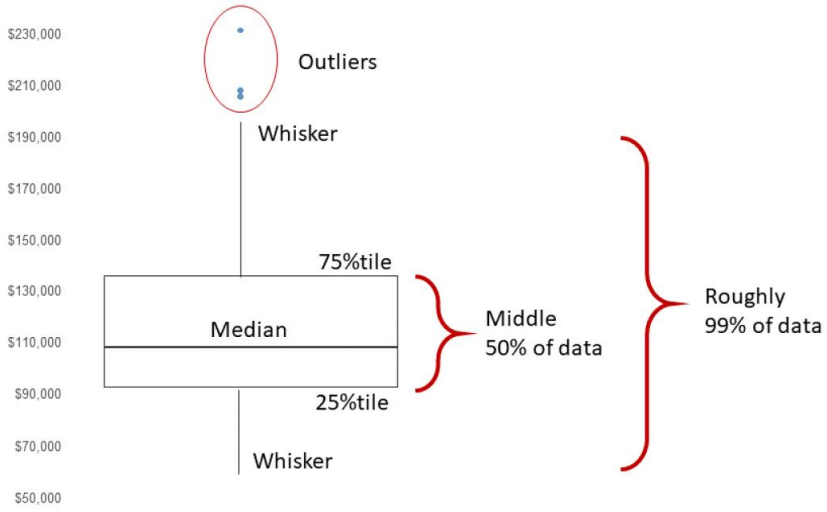


Grafik tersebut menjelaskan bahwa, secara umum, gaji naik sesuai dengan peringkat. Namun, untuk profesor memiliki kisaran gaji yang lebih luas.

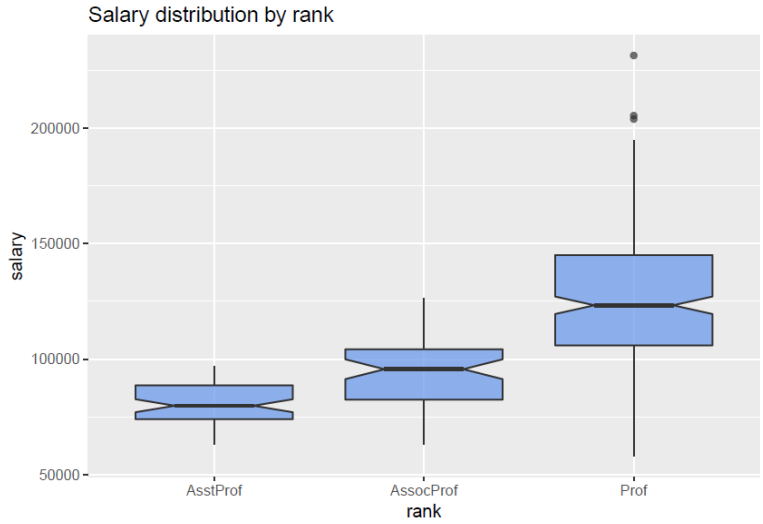
c. Box plot

Box plot dapat menampilkan persentil (25%), median, dan persentil (75%) dari suatu distribusi. Box plot berguna untuk membandingkan beberapa kelompok (yaitu, tingkat variabel kategori) pada variabel numerik.

```
# plot the distribution of salaries by rank using boxplots
ggplot(Salaries, aes(x = rank,
                     y = salary)) +
  geom_boxplot(notch = TRUE,
               fill = "cornflowerblue",
               alpha = .7) +
  labs(title = "Salary distribution by rank")
```



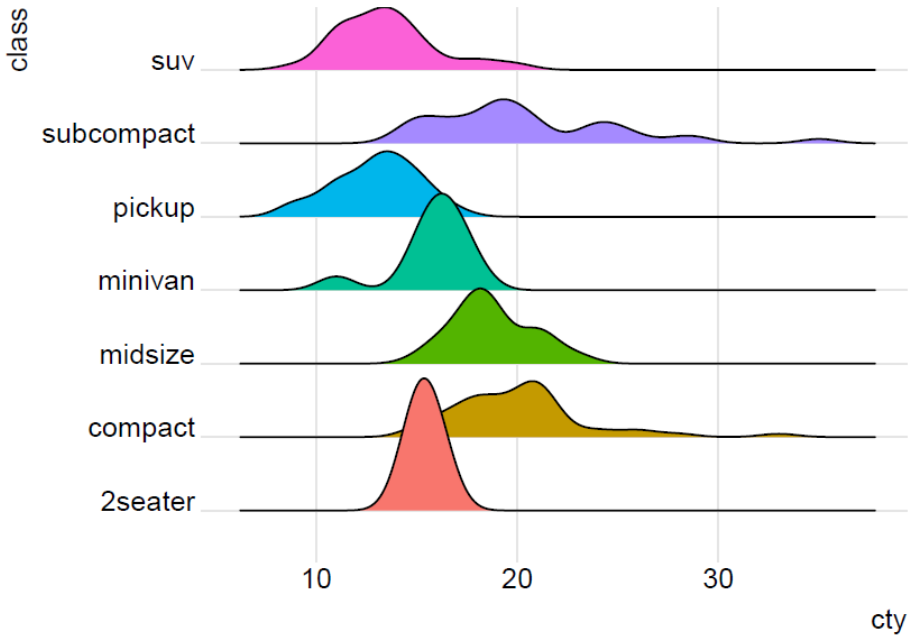
Output:



d. Ridgeline plots (joyplot)

Ridgeline plots mirip dengan kernel density plots yaitu digunakan untuk menampilkan distribusi variabel kuantitatif untuk beberapa kelompok. Namun, rifgeline plot menggunakan *faceting* vertikal sehingga tidak membutuhkan banyak ruang.

```
# create ridgeline graph
library(ggplot2)
library(ggridges)

ggplot(mpg,
       aes(x = cty,
           y = class,
           fill = class)) +
  geom_density_ridges() +
  theme_ridges() +
  labs("Highway mileage by auto class") +
  theme(legend.position = "none")
```



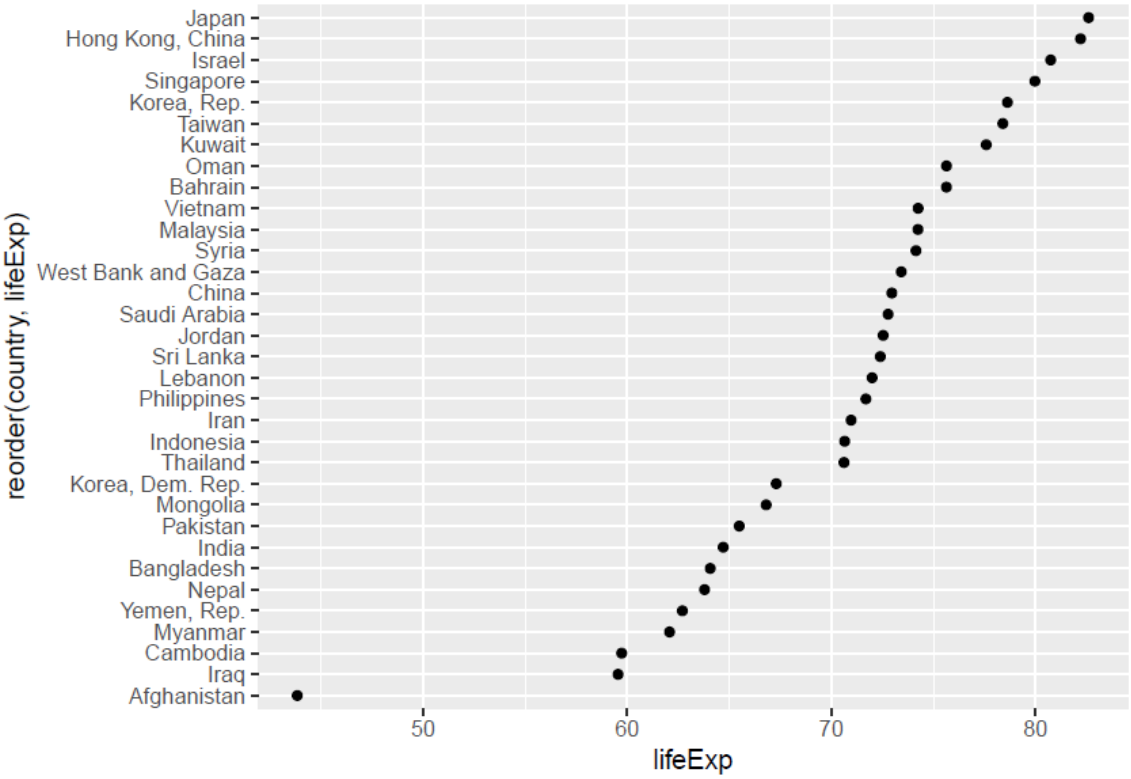
e. Cleveland Dot Charts

Cleveland Dot Charts berguna ketika anda ingin membandingkan statistik numerik untuk banyak grup. Sebagai contoh, anda ingin membandingkan harapan hidup tahun 2007 untuk negara Asia menggunakan dataset gapminder.

```
data(gapminder, package="gapminder")

# subset Asian countries in 2007
library(dplyr)
plotdata <- gapminder %>%
  filter(continent == "Asia" &
         year == 2007)

# basic Cleveland plot of life expectancy by country
ggplot(plotdata,
       aes(x= lifeExp, y = country)) +
  geom_point()
```

```
# Sorted Cleveland plot
ggplot(plotdata,
       aes(x=lifeExp,
           y=reorder(country, lifeExp))) +
  geom_point()
```

Output:



PERCOBAAN PRAKTIKUM:

1. Unduhlah dataset untuk kasus *house price prediction* di alamat berikut ini: https://www.kaggle.com/shree1992/housedata?select=output.csv
2. Pahami data tersebut, identifikasi jumlah dan tipe atributnya.
3. Visualisasikan menggunakan teknik visualisasi yang sudah dijelaskan pada modul ini.

TUGAS PRAKTIKUM:

Unduhlah (*download*) 1 dataset (bebas) dari Kaggle (https://www.kaggle.com/) atau UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php) kemudian identifikasi jenis atributnya dan visualisasikan menggunakan teknik visualisasi yang sudah dijelaskan pada modul ini.