Modern Data Science Methods for Educational Research

Creating Effective Data Visualization with R

Data Representation

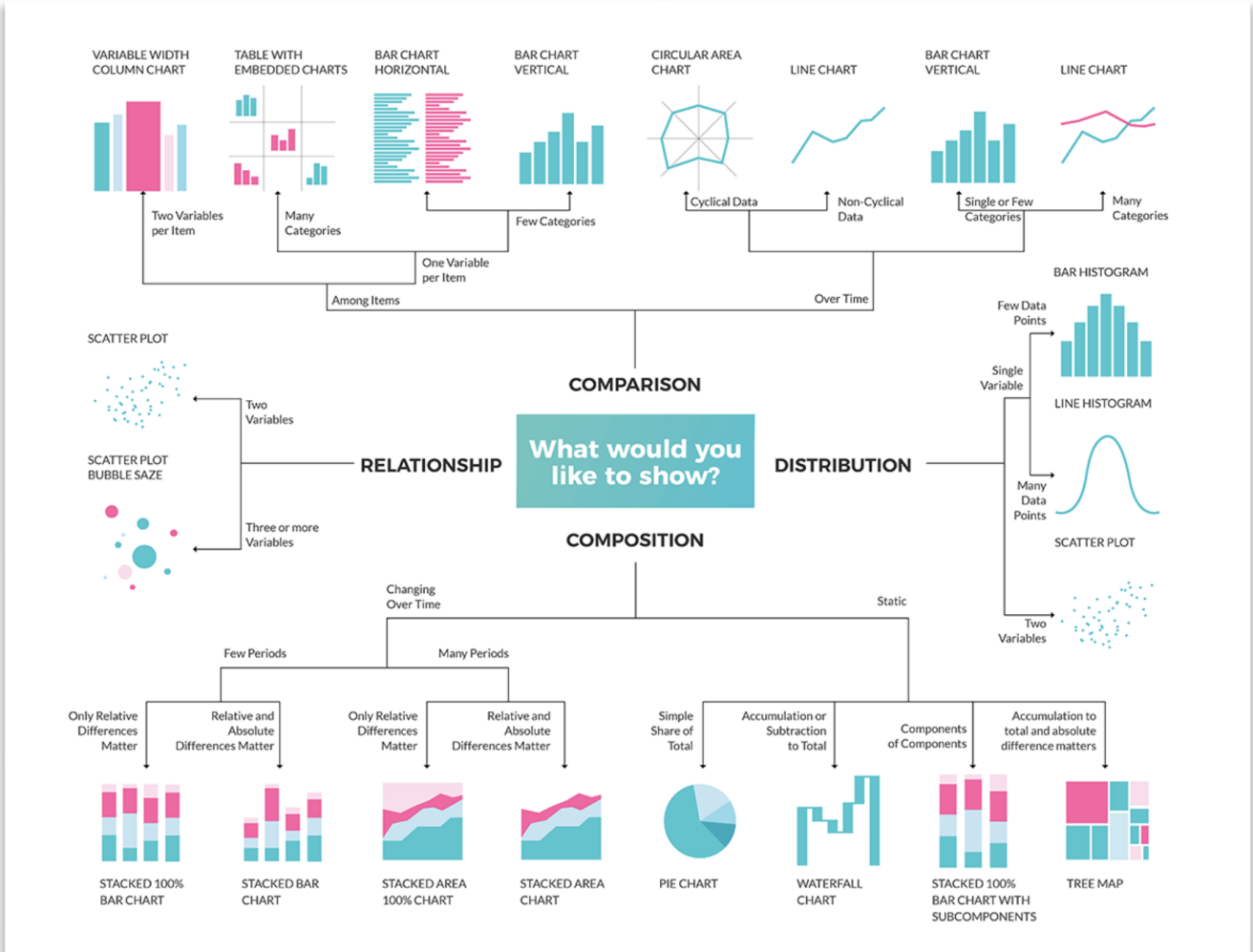
ผศ.ดร.สิวะโชติ ศรีสุทธิยากร

ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

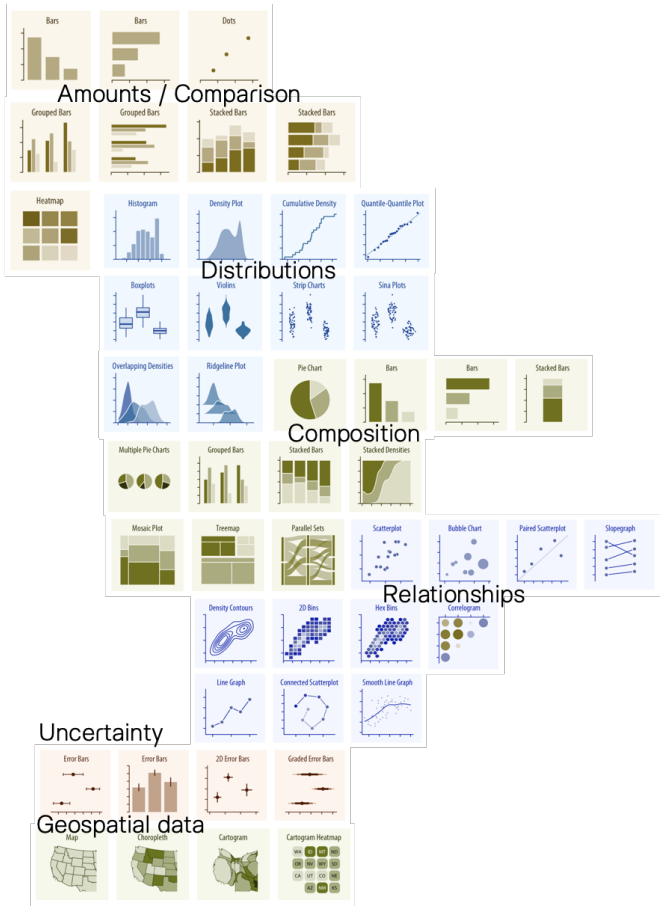
February 18, 2023

From Data to Visual

From Data to Visual



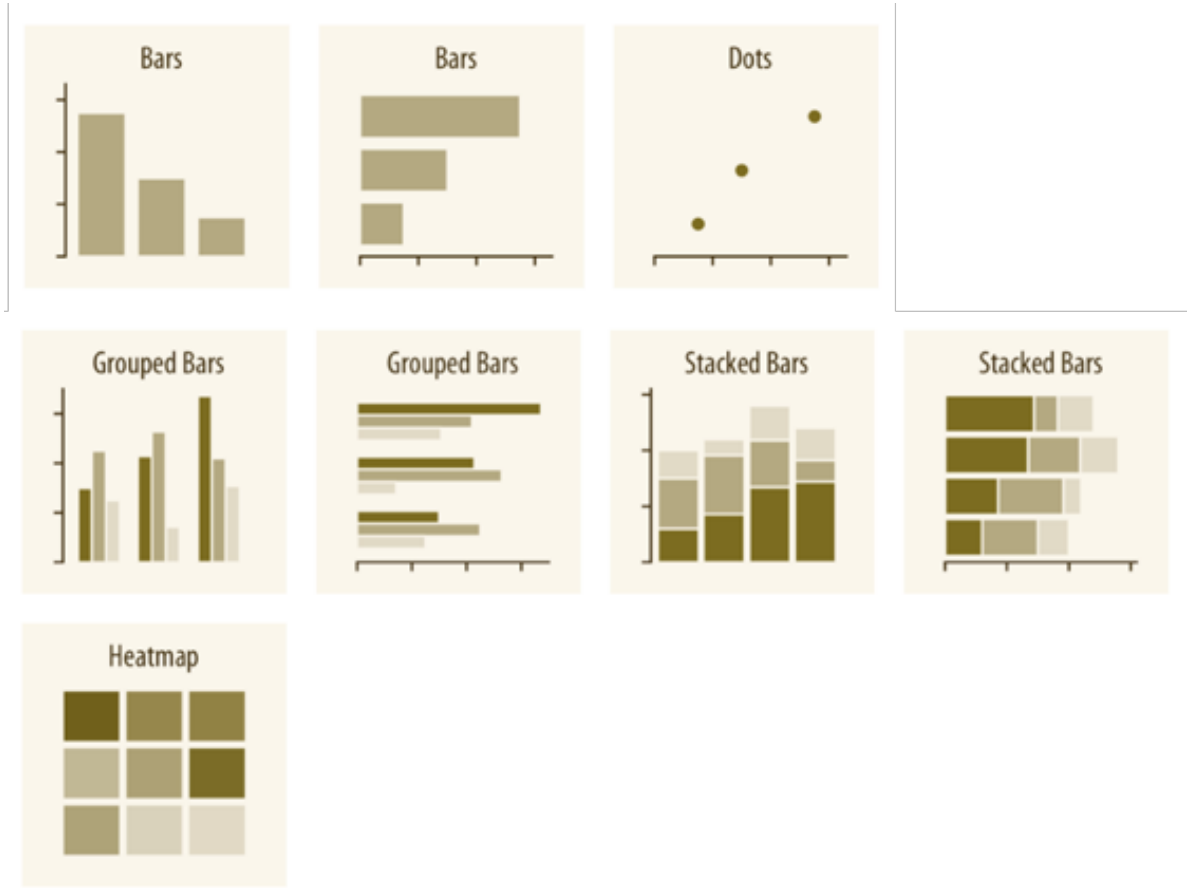
From Data to Visual



ที่มา : Wike (2019)

Visualizing Amounts

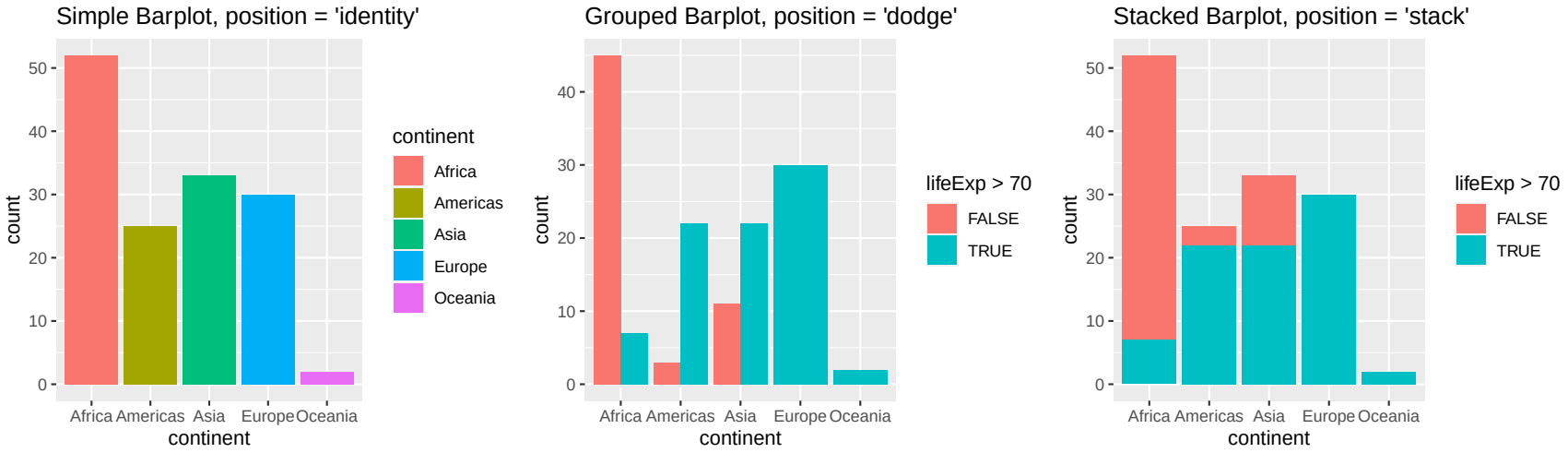
Visualizing Amounts



ที่มา : Wike (2019)

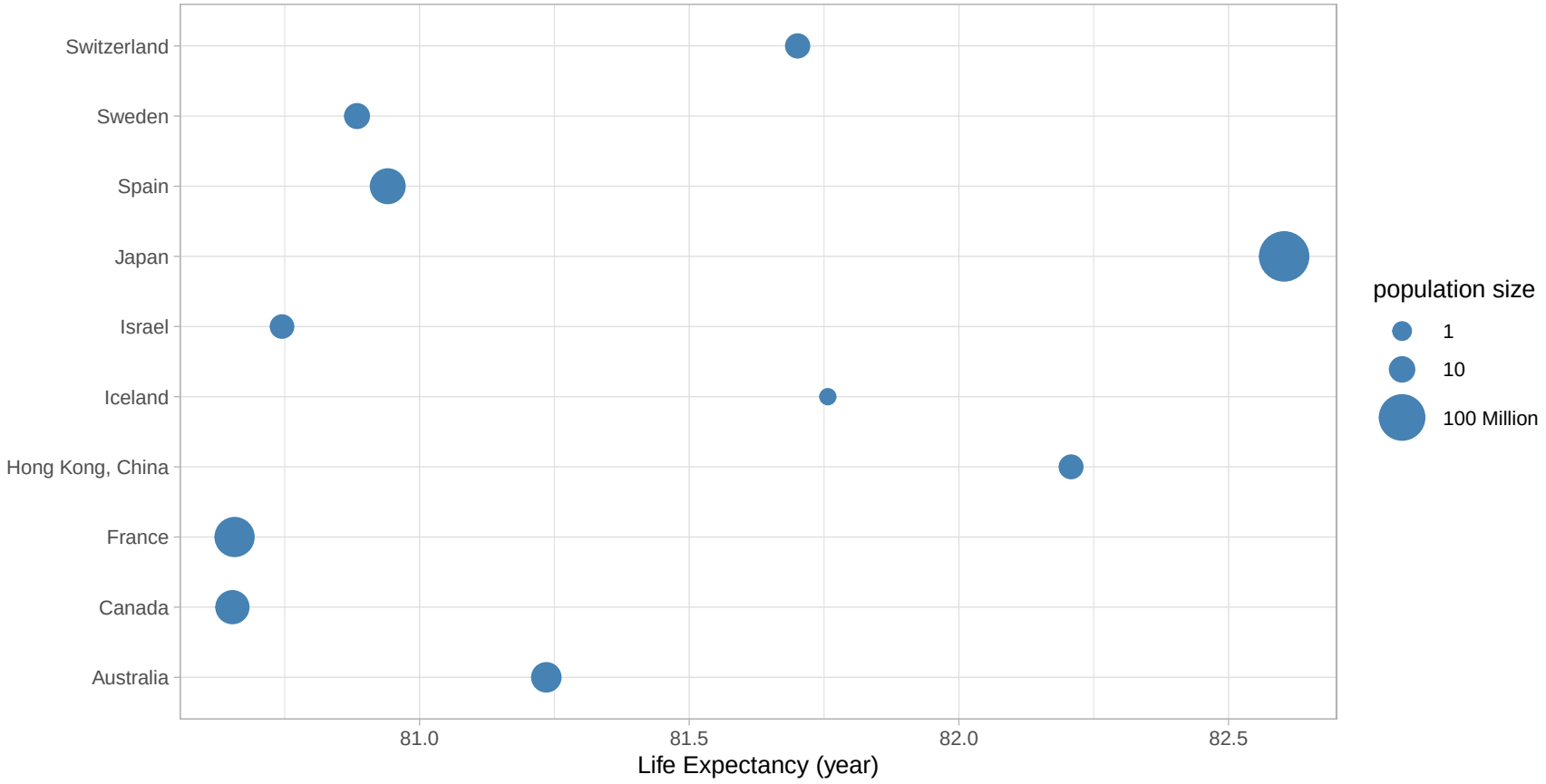
Various Barplots

`geom_bar()` ใช้สร้าง barplot ได้ทั้ง simple barplot, grouped barplot และ stacked barplot มีอาร์กิวเมนท์สำคัญได้แก่ `data`, `stat = "count"` หรือ `"identity"`, `position = "stack"` หรือ `"identity"` หรือ `"dodge"`, `width`, `orientation`



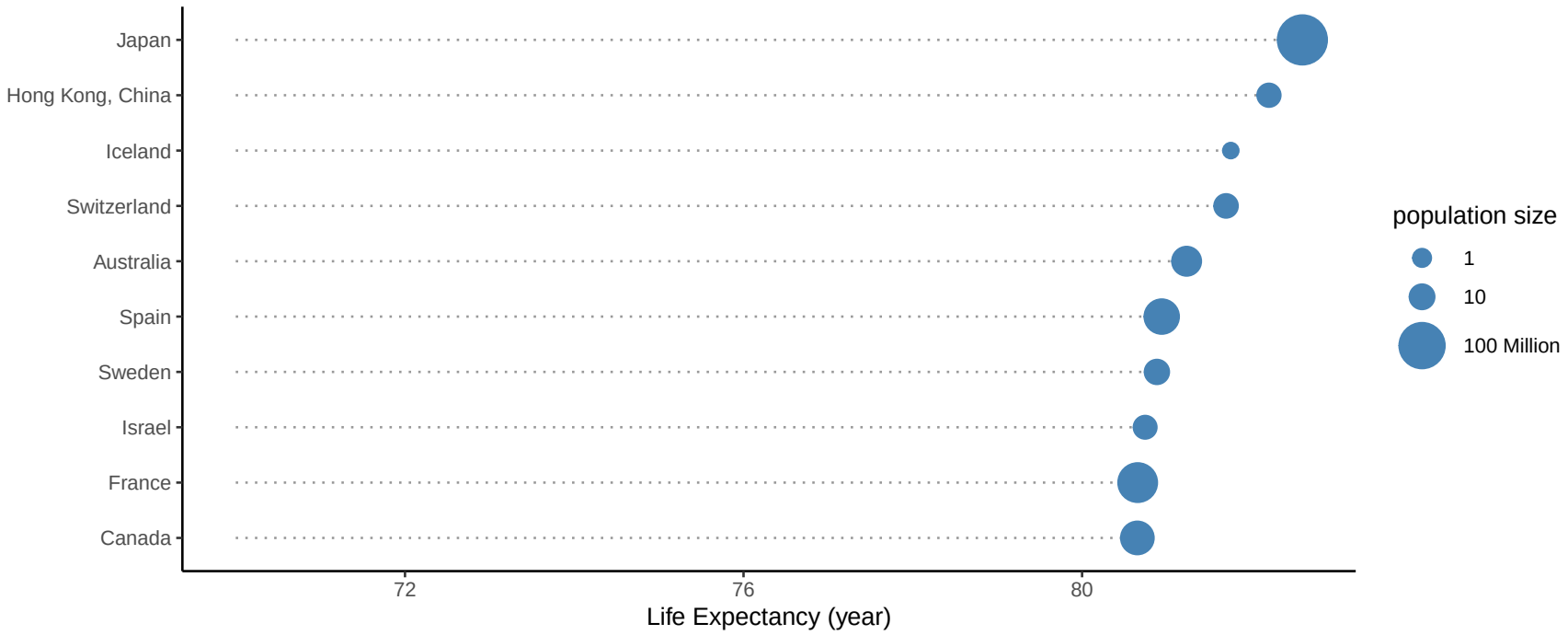
Dotplots

เหมือน barplot แต่เปลี่ยนจากการใช้ความยาวของแท่งสี่เหลี่ยมแทนข้อมูล เป็นการใช้จุดแทน



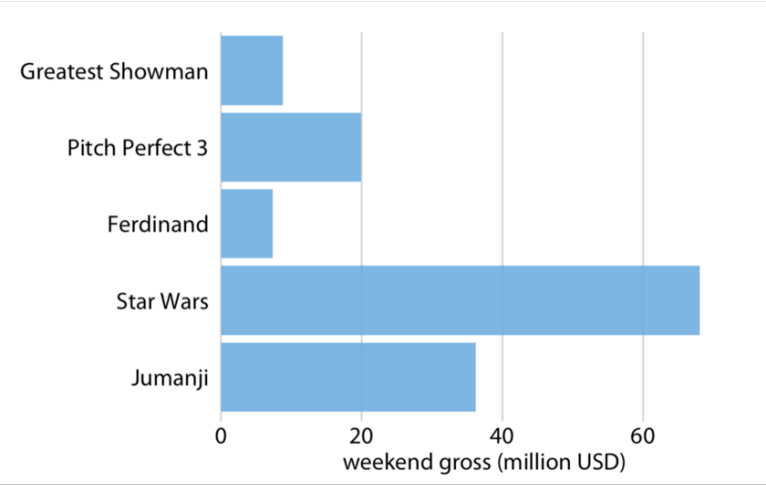
Lollipop plot

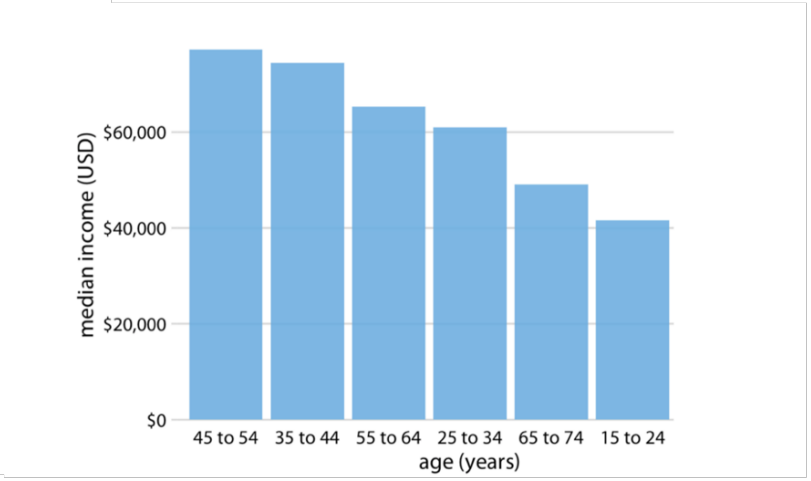
เหมือน dotplot แต่มีการใช้ segment ลากระหว่างแกนกับจุด เพื่อให้ชัดเจนเหมือน barplot มากขึ้น

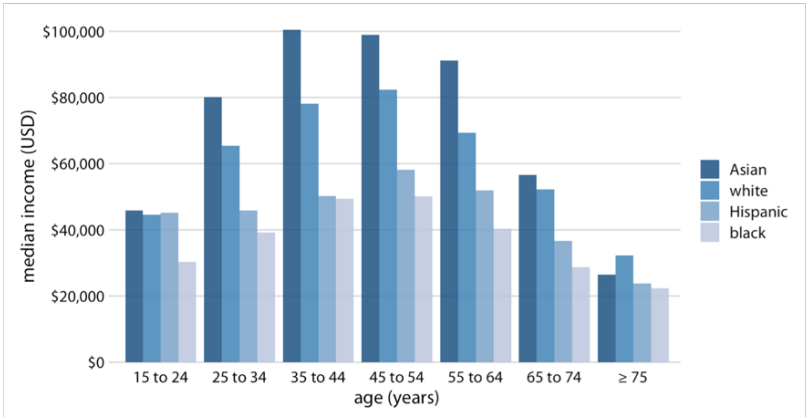


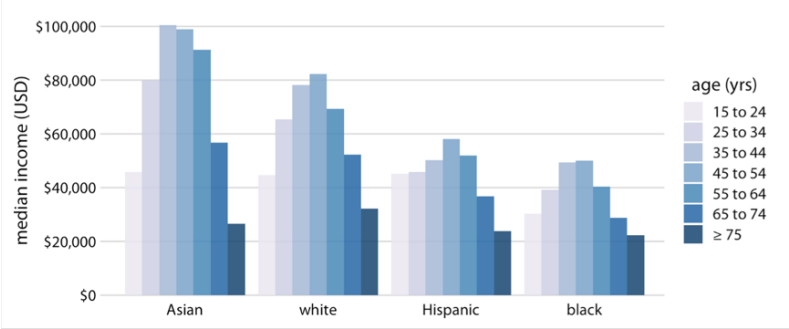
ข้อควรพิจารณาในการสร้าง barplot และ dotplot

Order is matter







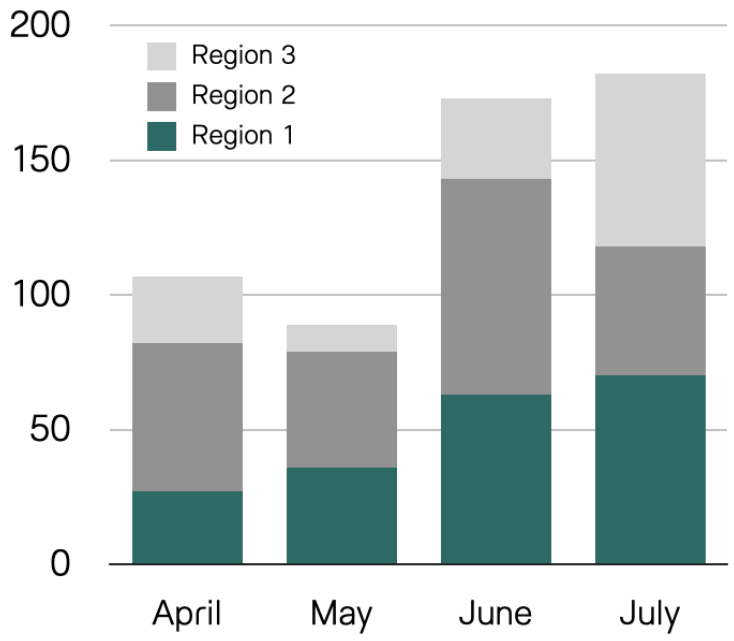


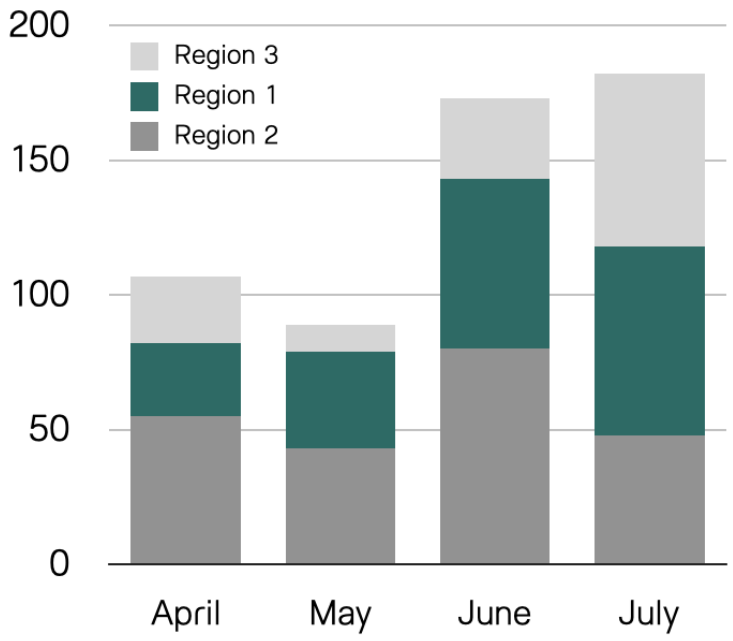
(Wilke, 2019)

ข้อควรพิจารณาในการสร้าง barplot และ dotplot

Order is matter

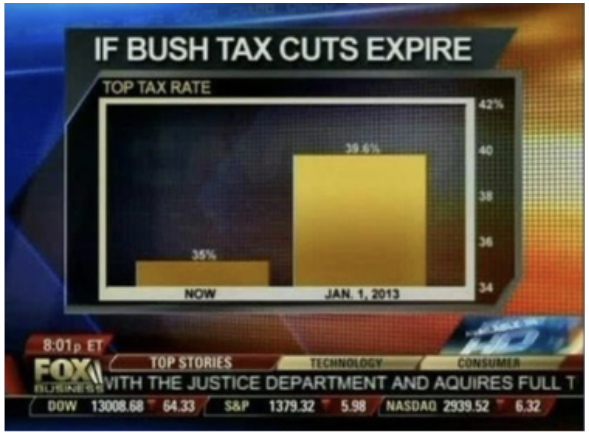
สำหรับ stacked bar plot การเปรียบเทียบความแตกต่างระหว่าง category ที่ไม่ได้เริ่มจากเส้น y=0 ทำได้ยาก

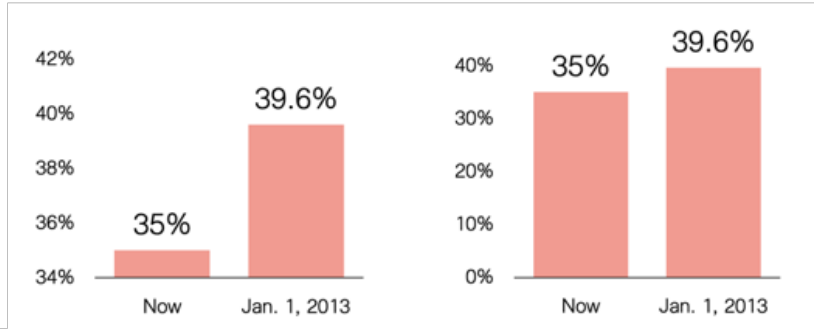


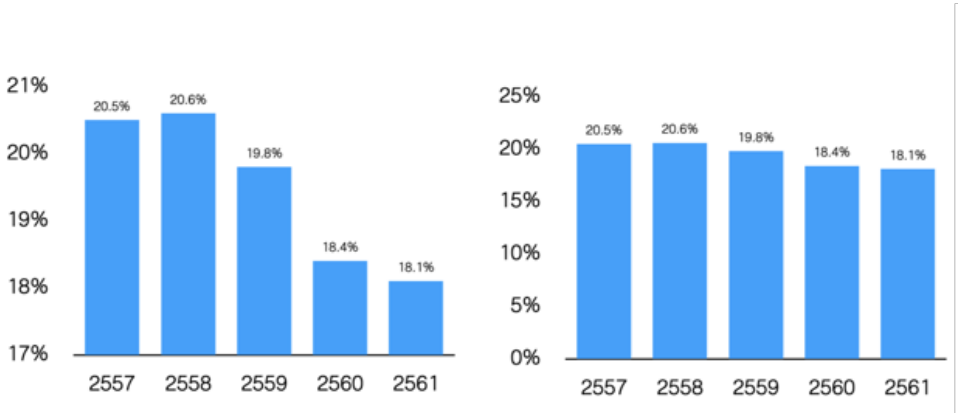


ข้อควรพิจารณาในการสร้าง barplot และ dotplot

อย่าลืมว่า plot ในกลุ่มนี้ ใช้ความยาวแทนข้อมูล







ข้อควรพิจารณาในการสร้าง barplot และ dotplot

อย่าลืมว่า plot ในกลุ่มนี้ ใช้ความยาวแทนข้อมูล

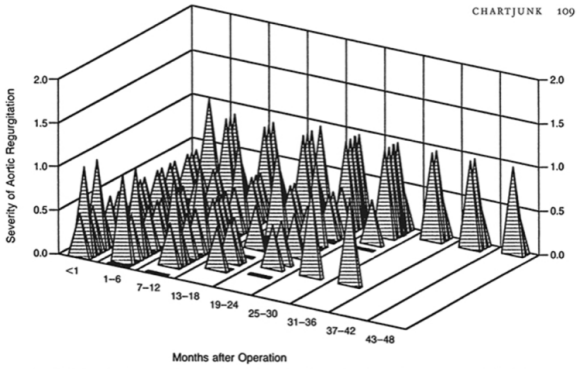
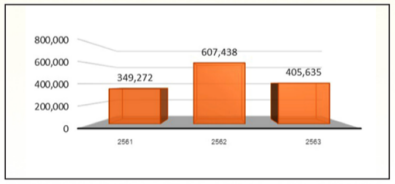
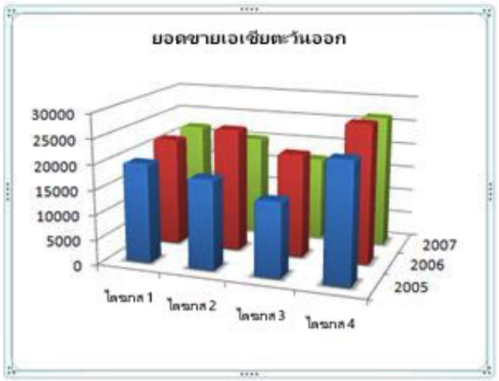


Figure 2. Serial Echocardiographic Assessments of the Severity of Regurgitation in the Pulmonary Autograft in 31 Patients. The numerical grades were assigned according to the severity of regurgitation, as follows: 0, none; 0.5, trivial; 1.0 to 1.5, mild; 2.0, moderate; and 3.0, severe.

แผนภาพที่ 4.5 จำนวนนักเรียนพิการเรียนร่วมในโรงเรียนปกติที่สังกัดสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (ปีการศึกษา 2561-2563)

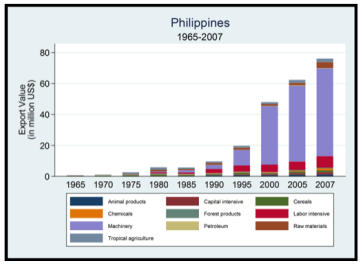


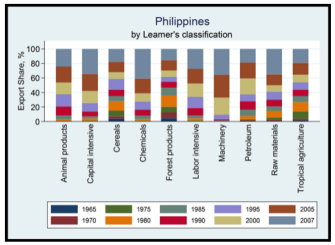
ที่มา: สำนักนโยบายและแผนการศึกษาขั้นพื้นฐาน สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน

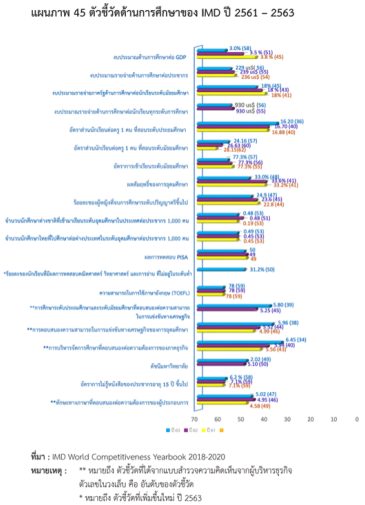


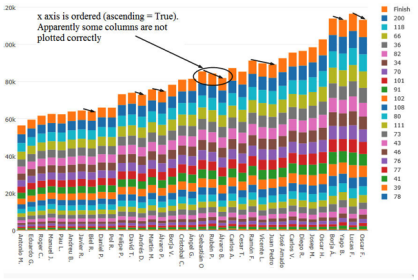
อย่าสร้างภาระให้กับผู้รับสาร !!

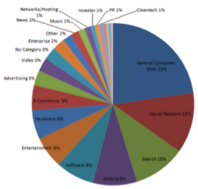
stacked bar plot (บางครั้งก็ grouped bar plot)
อาจสร้างภาระให้กับผู้รับสารได้ง่าย







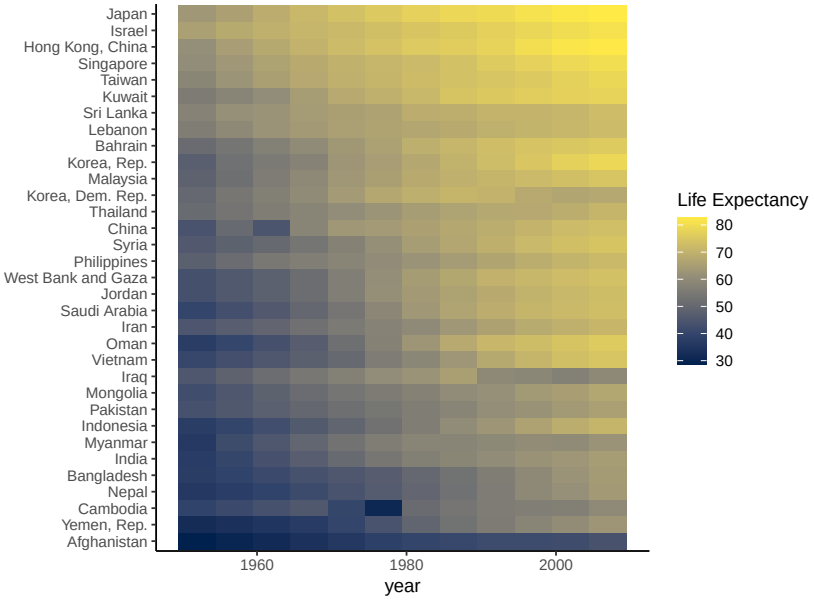


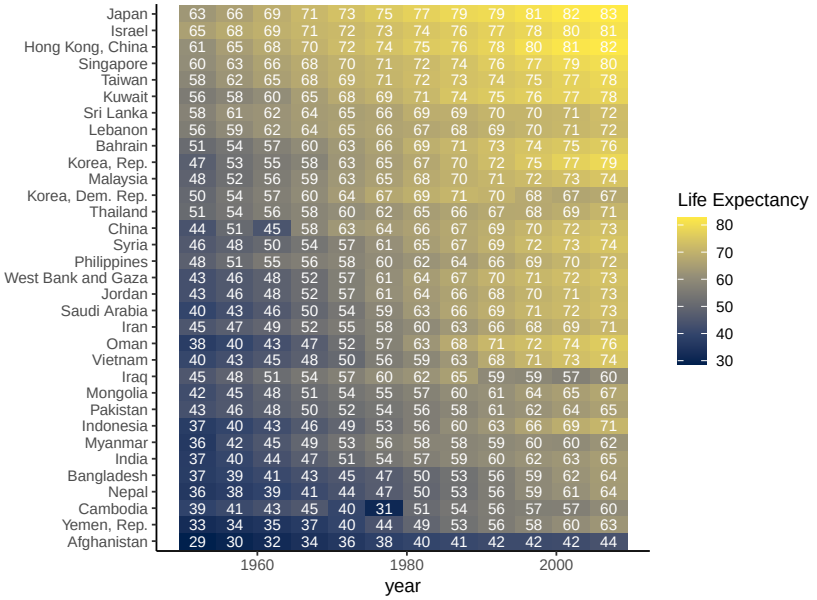


Heatmap

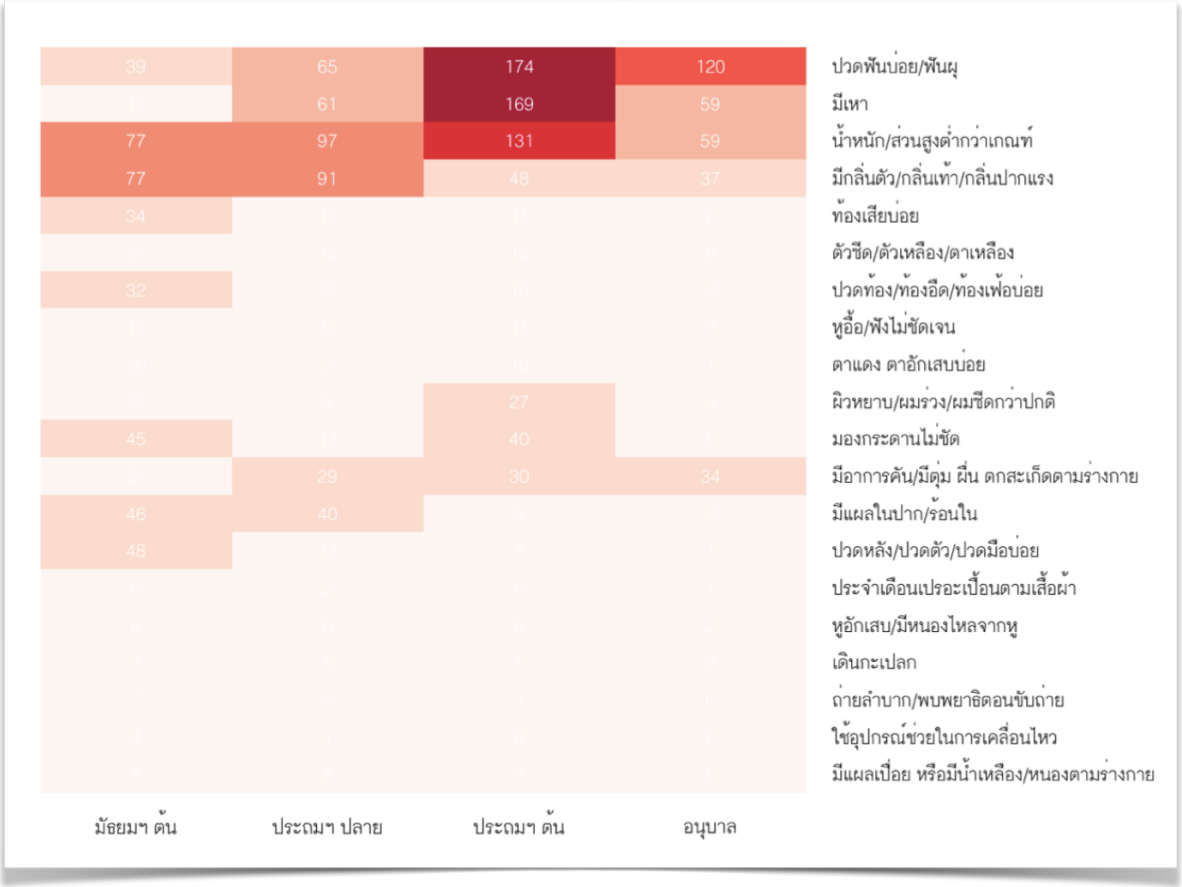
heatmap เหมือนกับตารางข้อมูล ที่มีการใช้สีแทนปริมาณข้อมูล

สามารถใช้ได้กับทั้งข้อมูล cross-sectional data และ time-series data



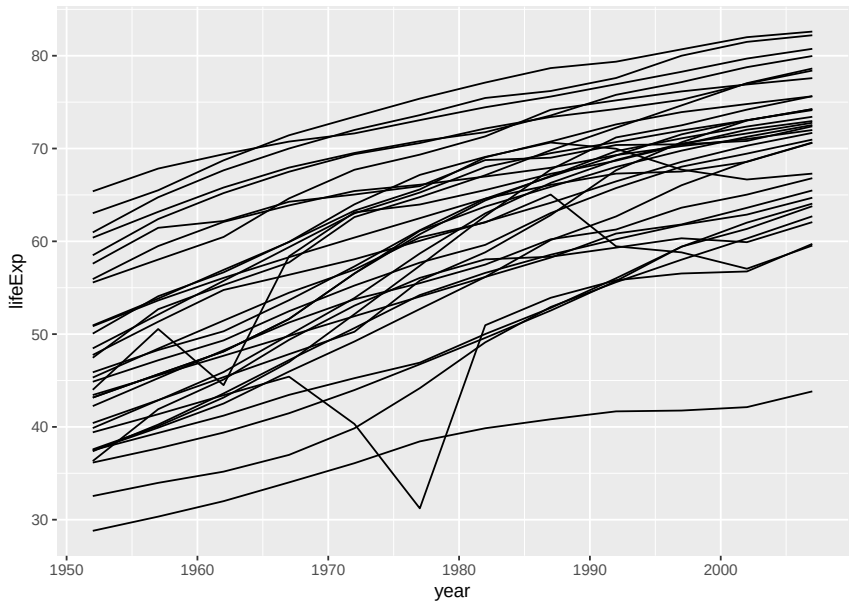


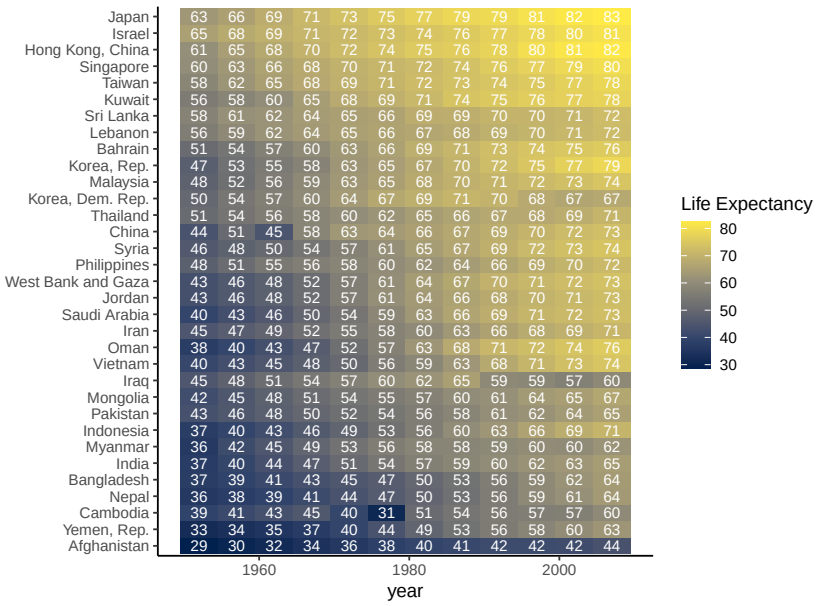
Heatmap


39 65 174 120 ปวดฟันบ่อย/ฟันผุ
16 61 169 59 มีเหา
77 97 131 59 น้ำหนัก/ส่วนสูงต่ำกว่าเกณฑ์
77 91 48 37 มีกลิ่นตัว/กลิ่นเท้า/กลิ่นปากแรง
34 10 11 6 ท้องเสียบ่อย
21 15 12 8 ตัวซีด/ตัวเหลือง/ตาเหลือง
32 22 19 13 ปวดท้อง/ท้องอืด/ท้องเฟ้อบ่อย
19 6 11 3 หูอื้อ/ฟังไม่ชัดเจน
20 2 10 1 ตาแดง ตาอักเสบบ่อย
22 5 27 3 ผิวหยาบ/ผมร่วง/ผมซีดกว่าปกติ
45 18 40 5 มองกระดานไม่ชัด
21 29 30 34 มีอาการคัน/มีตุ่ม ผื่น ตกสะเก็ดตามร่างกาย
46 40 9 17 มีแผลในปาก/ร้อนใน
48 19 7 1 ปวดหลัง/ปวดตัว/ปวดมือบ่อย
11 2 6 1 ประจำเดือนเปรอะเปื้อนตามเสื้อผ้า
6 0 6 2 หูอักเสบ/มีหนองไหลจากหู
7 5 5 3 เดินกะเปลก
7 5 1 6 ถ่ายลำบาก/พบพยาธิตอนขับถ่าย
4 1 0 0 ใช้อุปกรณ์ช่วยในการเคลื่อนไหว
9 8 7 6 มีแผลเปื่อย หรือมีน้ำเหลือง/หนองตามร่างกาย
มัธยมฯ ต้น
ประถมฯ ปลาย
ประถมฯ ต้น
อนุบาล

Heatmap

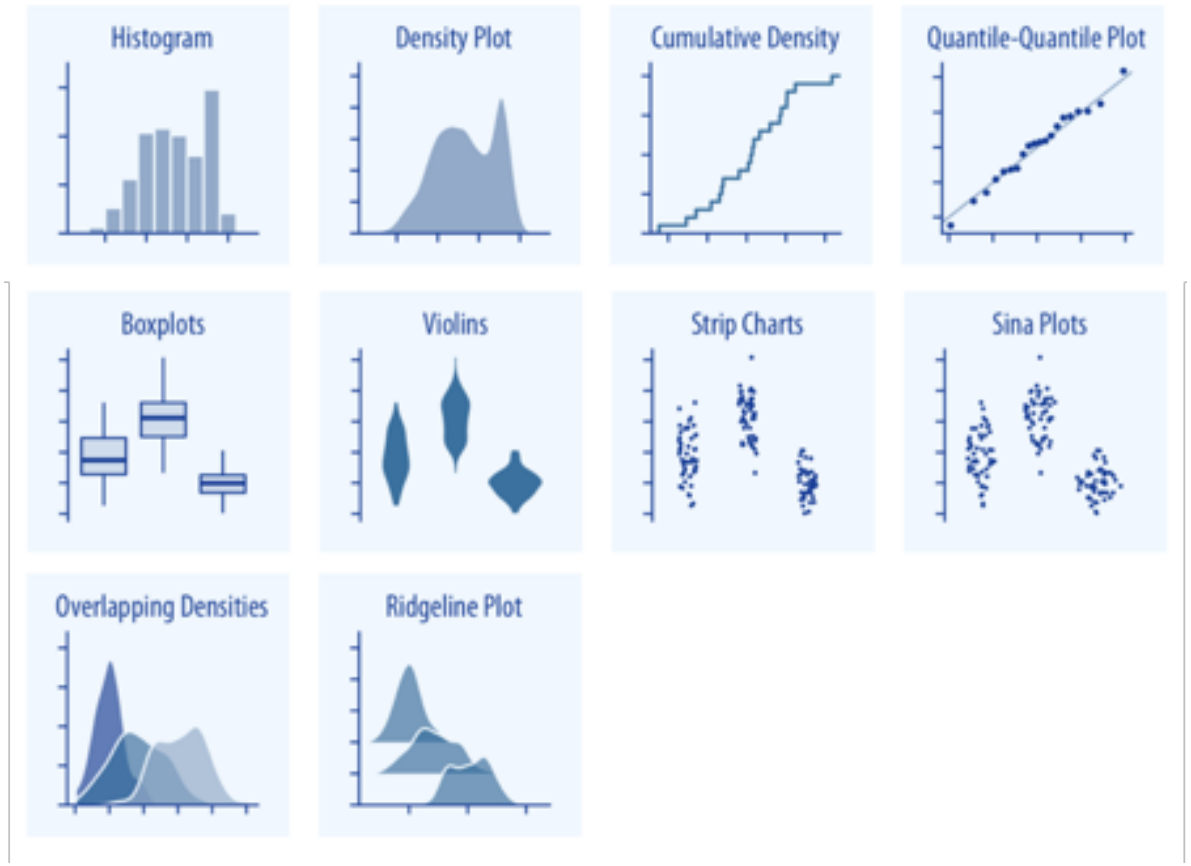
heatmap มีประสิทธิภาพในการนำเสนอข้อมูล สูงกว่า line plot ในกรณีที่มีหน่วยข้อมูลที่ต้องการศึกษาแนวโน้มจำนวนมาก





Visualizing Distribution

Visualizing Distribution

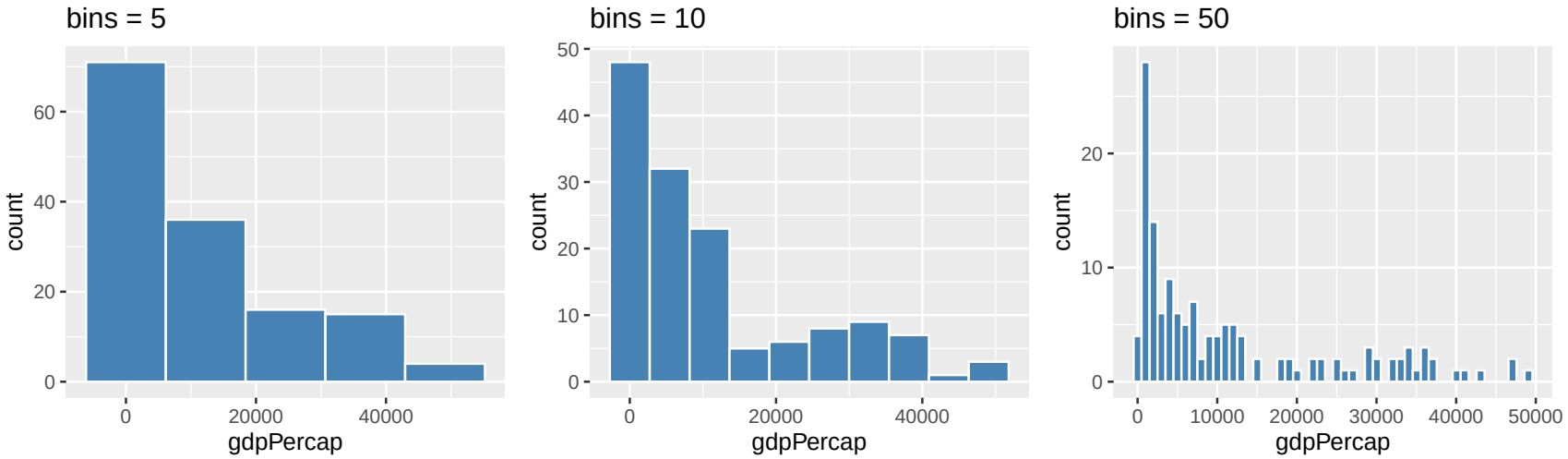


ที่มา : Wike (2019)

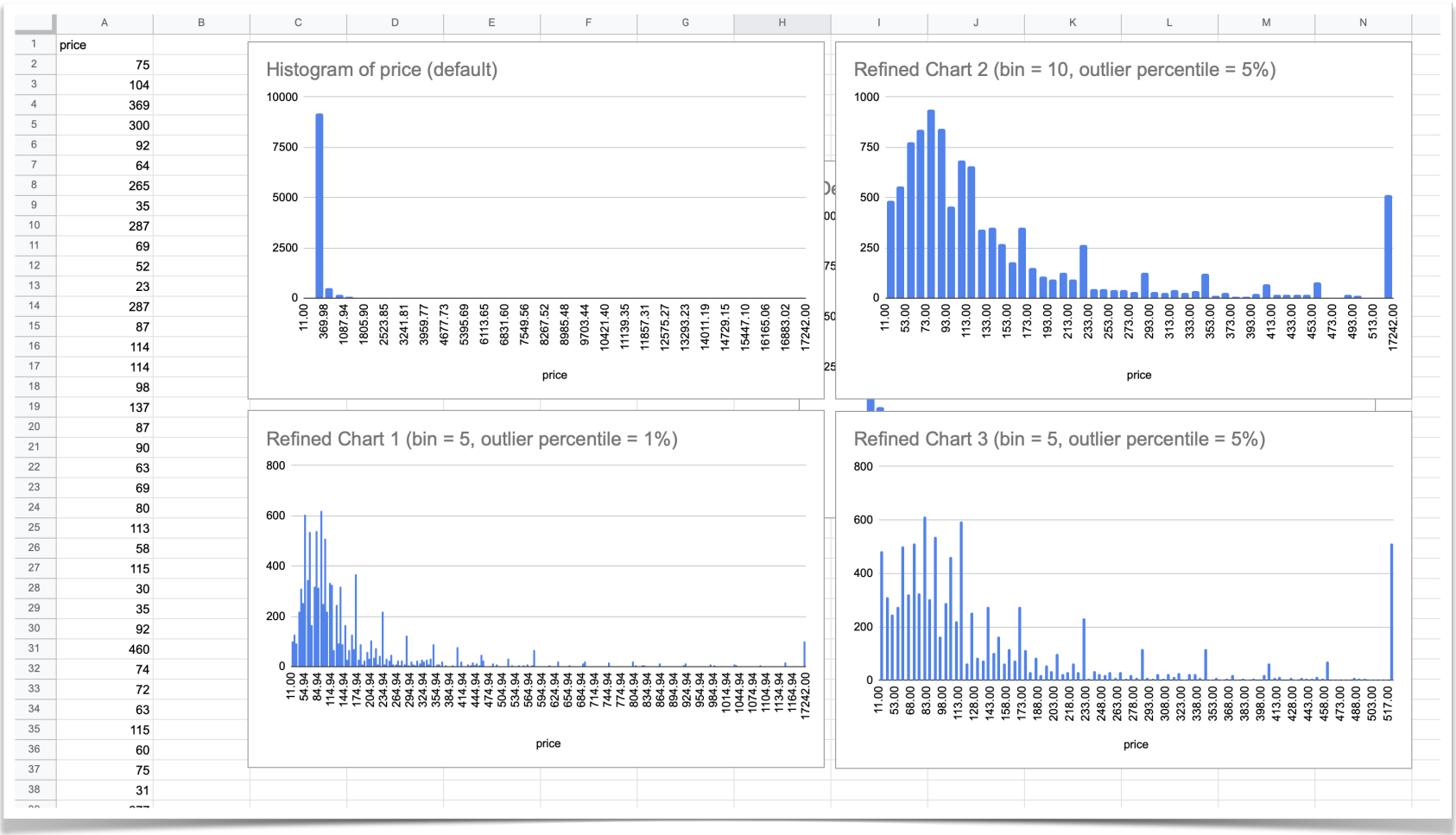
Single Variable: Histogram

histogram สร้างจากตารางแจกแจงความถี่แบบมีอันตรภาคชั้น
ความกว้างของอันตรภาคชั้น (bins) จึงเป็นพารามิเตอร์ตัวหนึ่งของ histogram

```
gapminder %>%
  filter(year == "2007") %>%
  ggplot(aes(x = gdpPercap))+
  geom_histogram(fill = "steelblue", col="white")
```



Histogram: binwidth



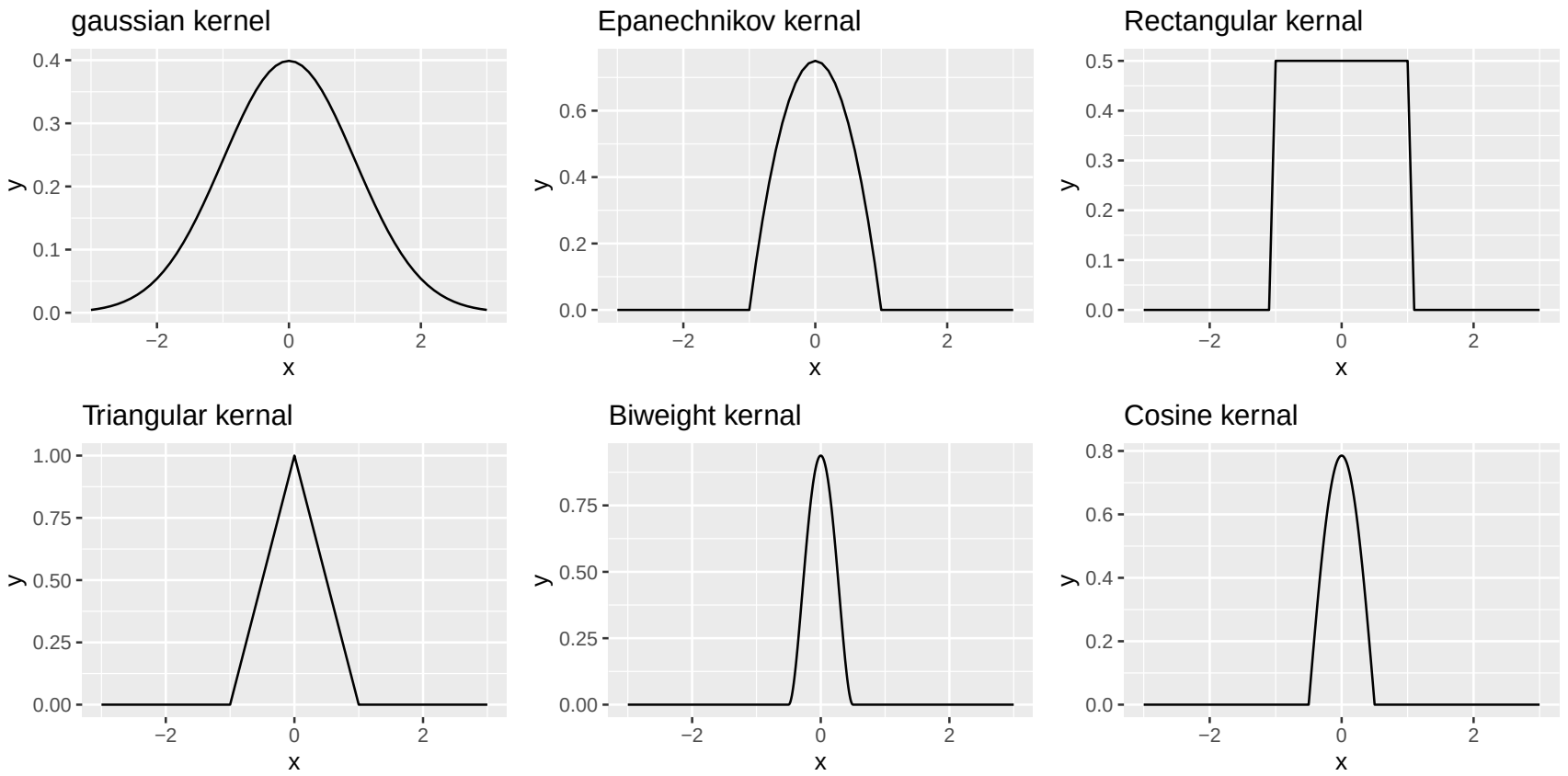
Single Variable: Density plot

แผนภาพโค้งความหนาแน่น (density plot)
ใช้ประมาณการแจกแจงความน่าจะเป็นเบื้องหลังของตัวแปร โดยทั่วไป
จะใช้อัลกอริทึมที่เรียกว่า kernal density estimation (KDE)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

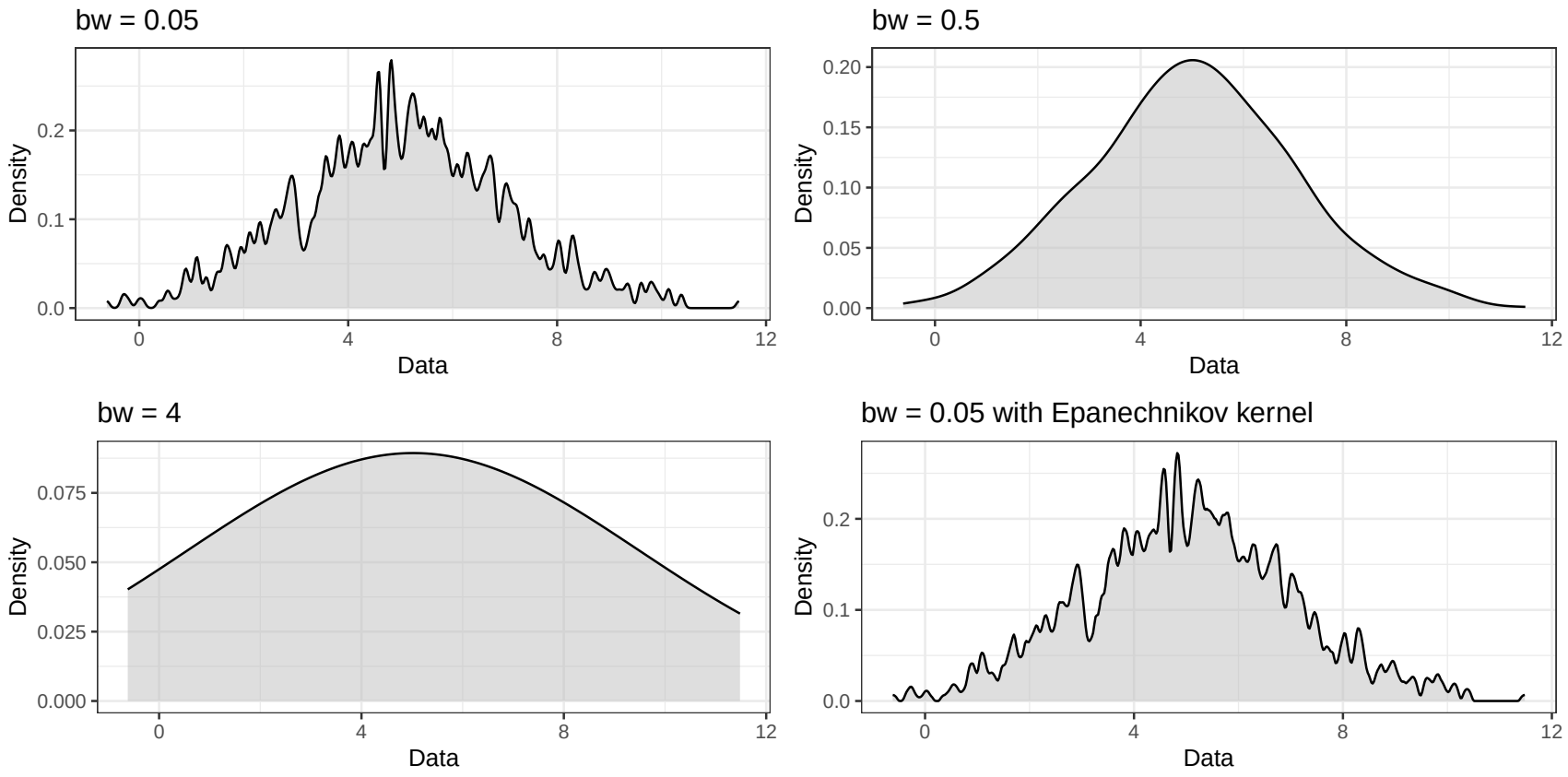
Density plot: Kernel Functions

kernel function เป็นพารามิเตอร์ตัวหนึ่งของ density plot

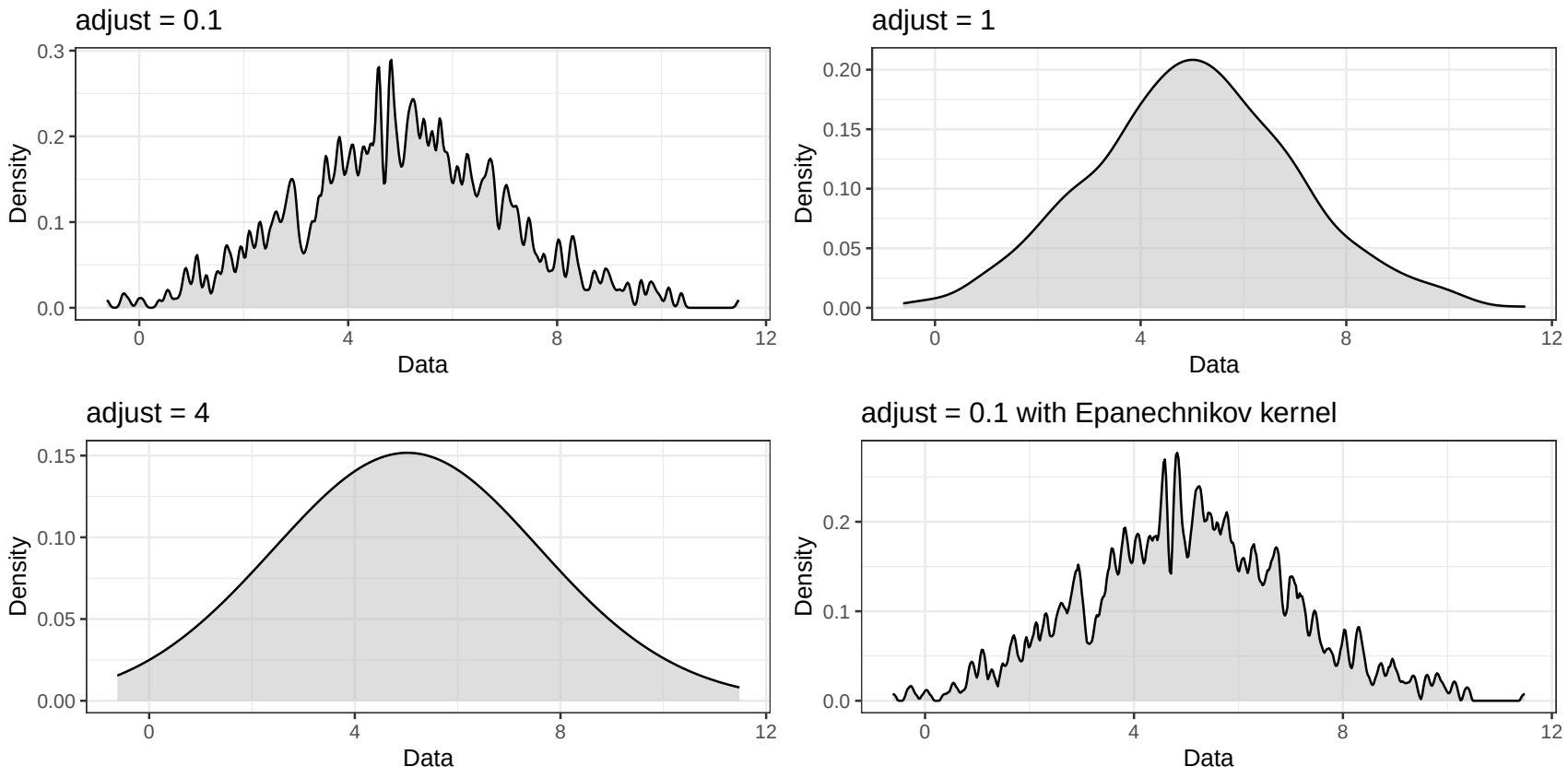


Density plot: bandwidth

bandwidth เป็นพารามิเตอร์อีกตัวหนึ่งของ density plot ซึ่งสามารถกำหนดได้จากอาร์กิวเมนท์ `bw` หรือ `adjust`

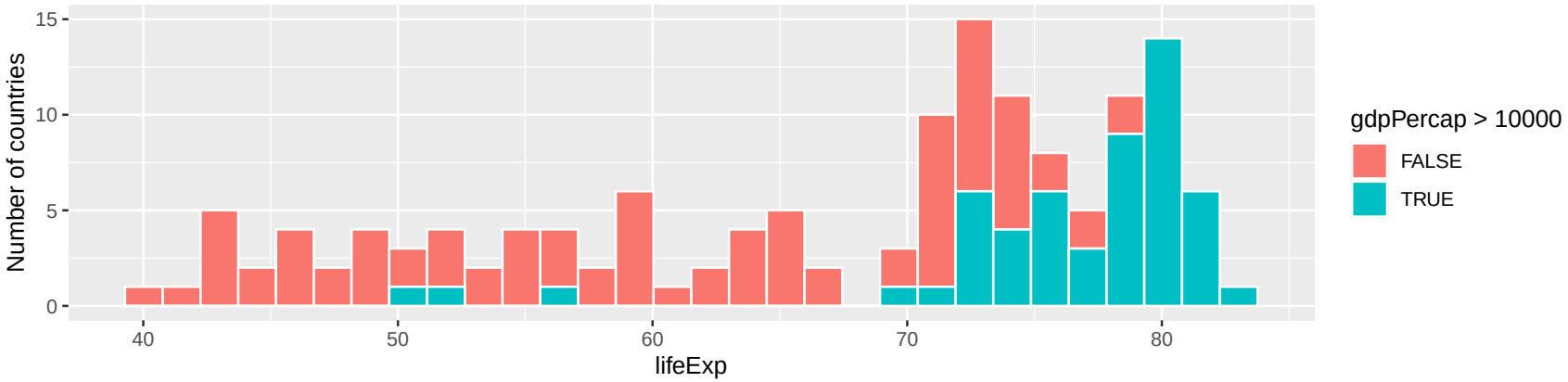


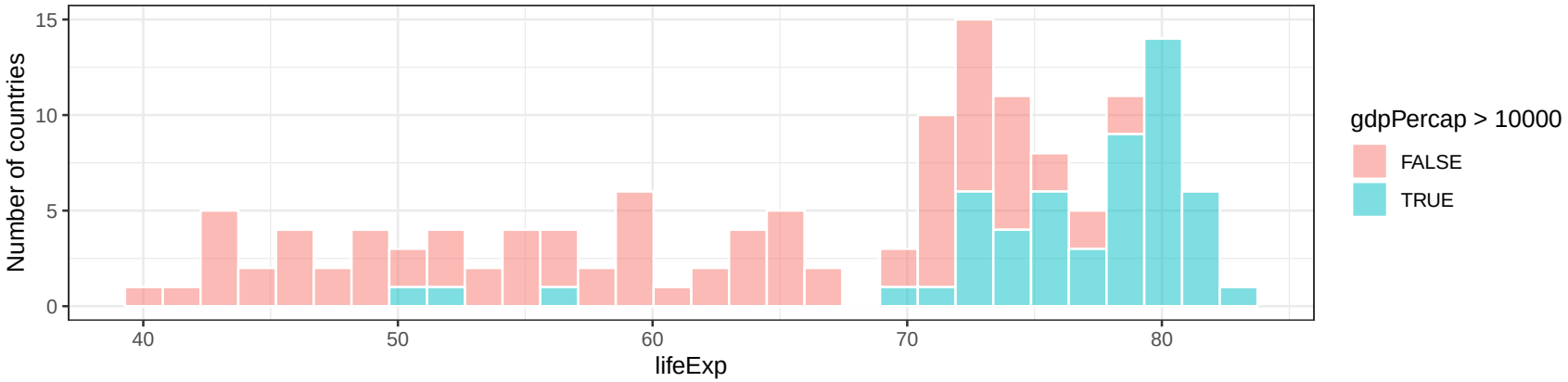
Density plot: bandwidth



Multiple Distributions: histograms

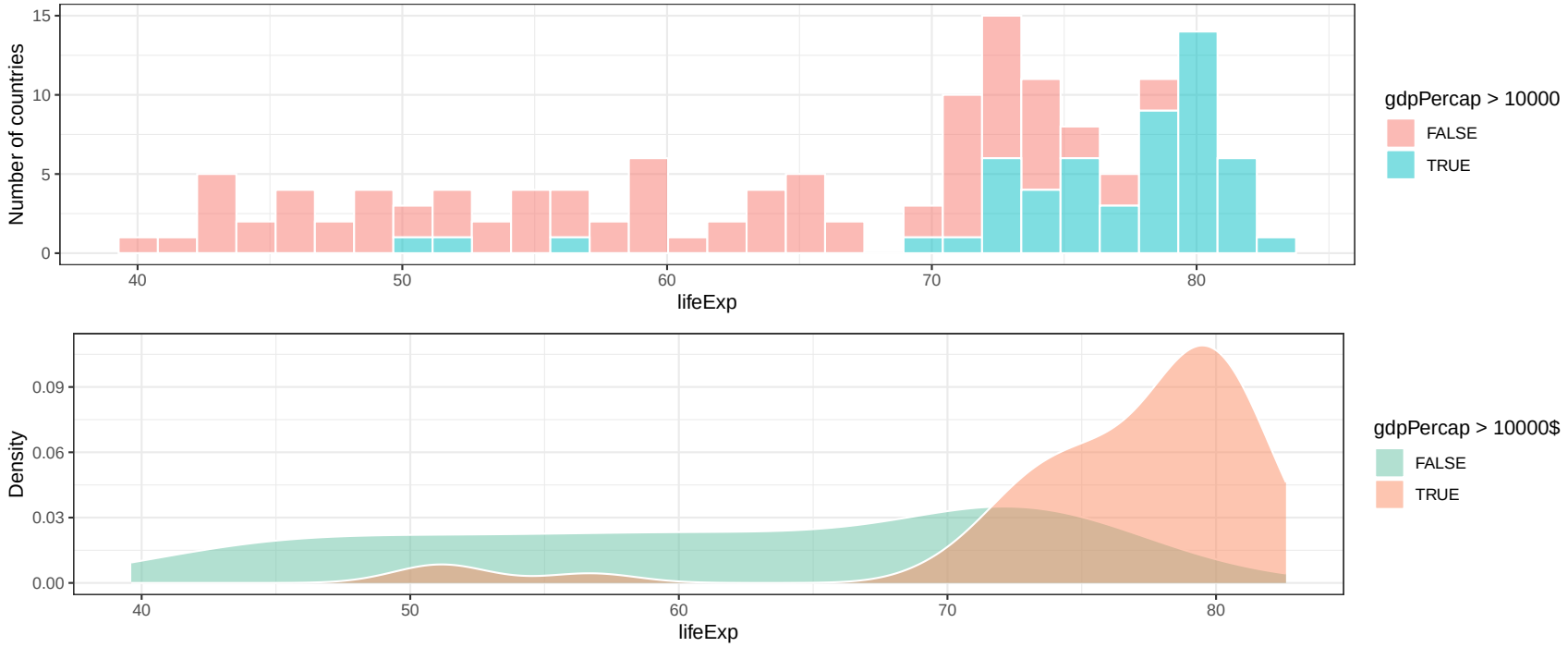
Good visualization ?





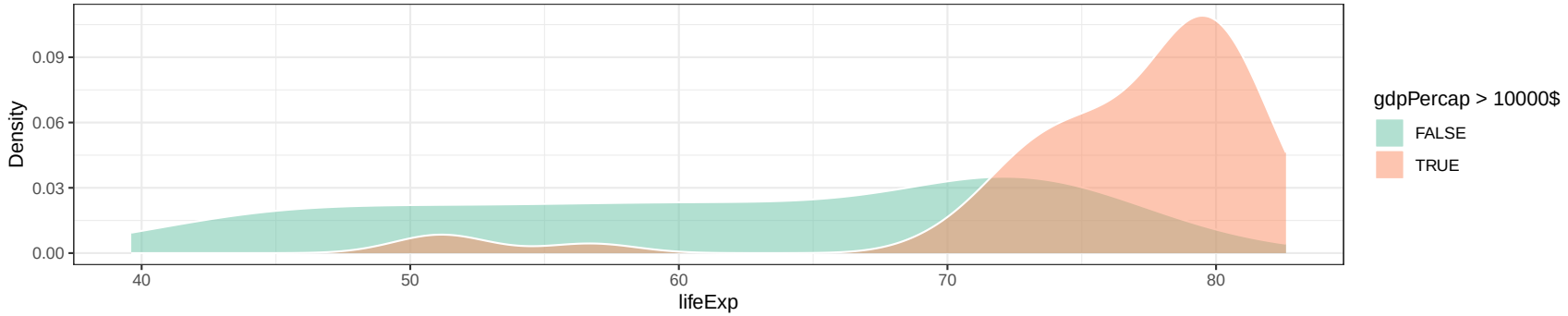
Multiple Distributions: densities

แผนภาพไหนเหมาะสมมากกว่ากัน ?

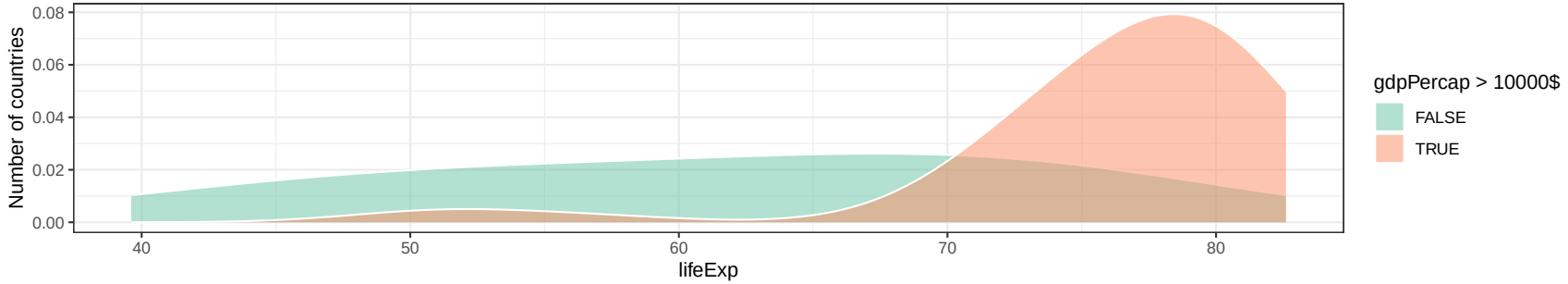


Multiple Distributions: densities

แผนภาพไหนเหมาะสมมากกว่ากัน ?



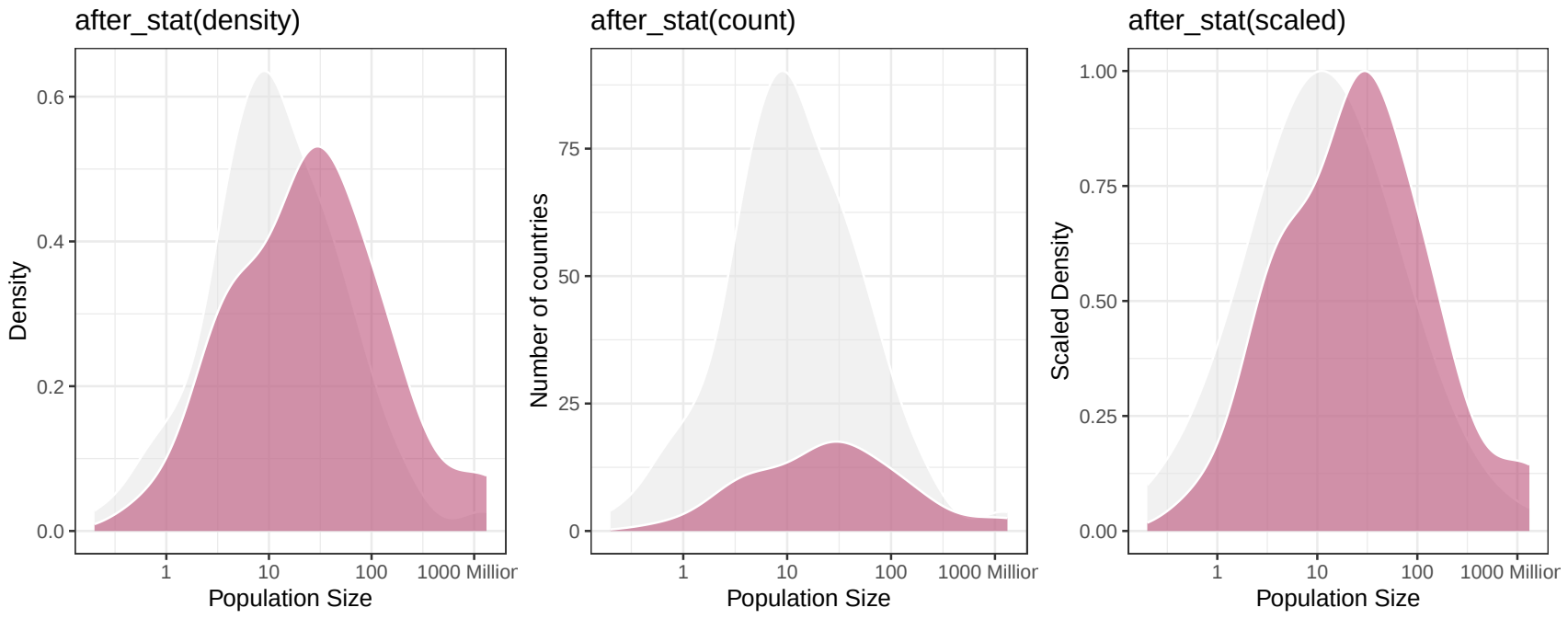
use adjust = 2



Multiple Distributions: densities

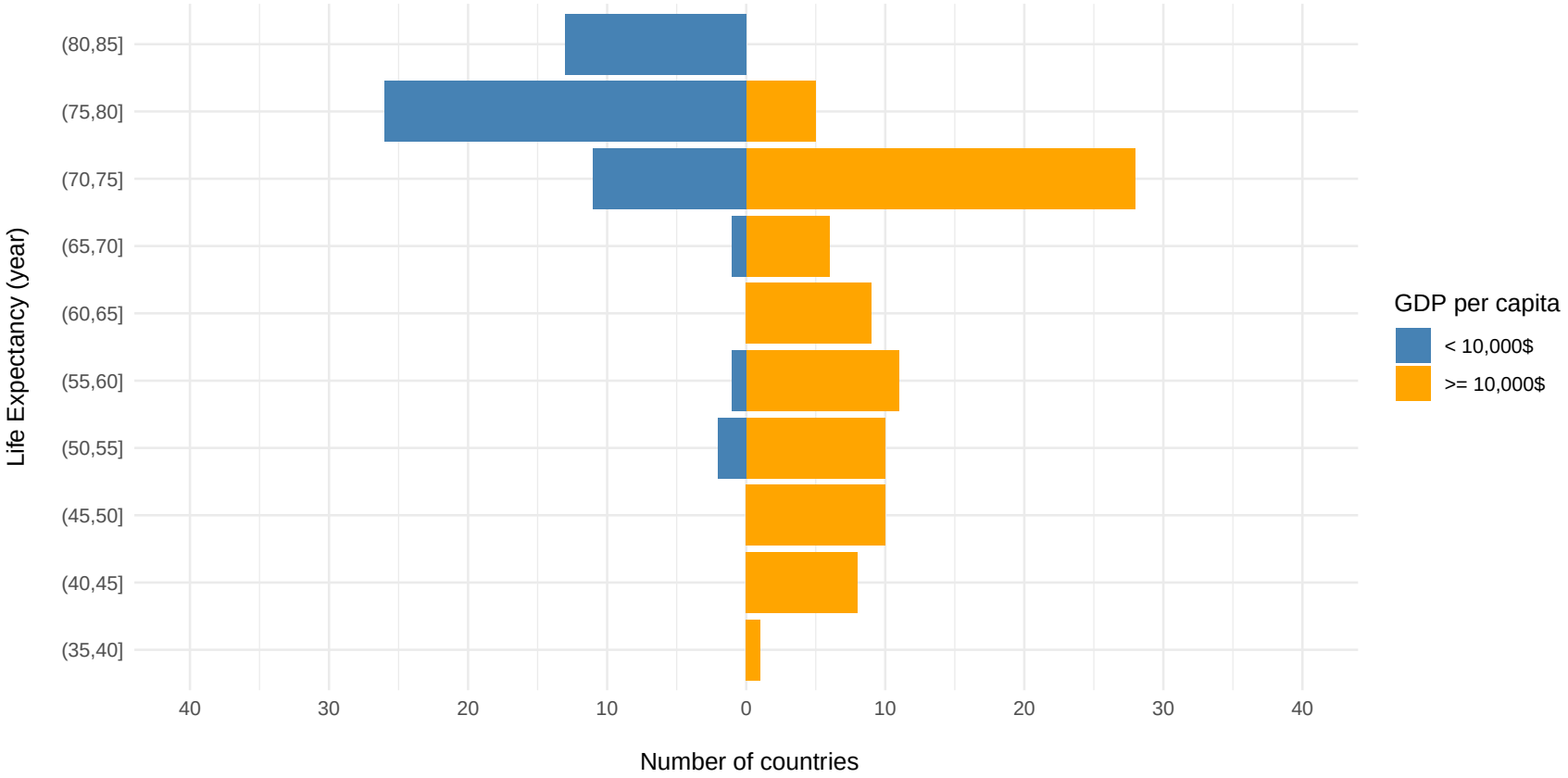
multiple densities สามารถใช้นำเสนอสารสนเทศในเชิงเปรียบเทียบ ส่วนย่อยเมื่อเปรียบเทียบกับส่วนทั้งหมด (part of whole) ได้ด้วย

จำนวนประชากรของประเทศในทวีปเอเชียเปรียบเทียบกับประเทศทั้งหมดบนโลก

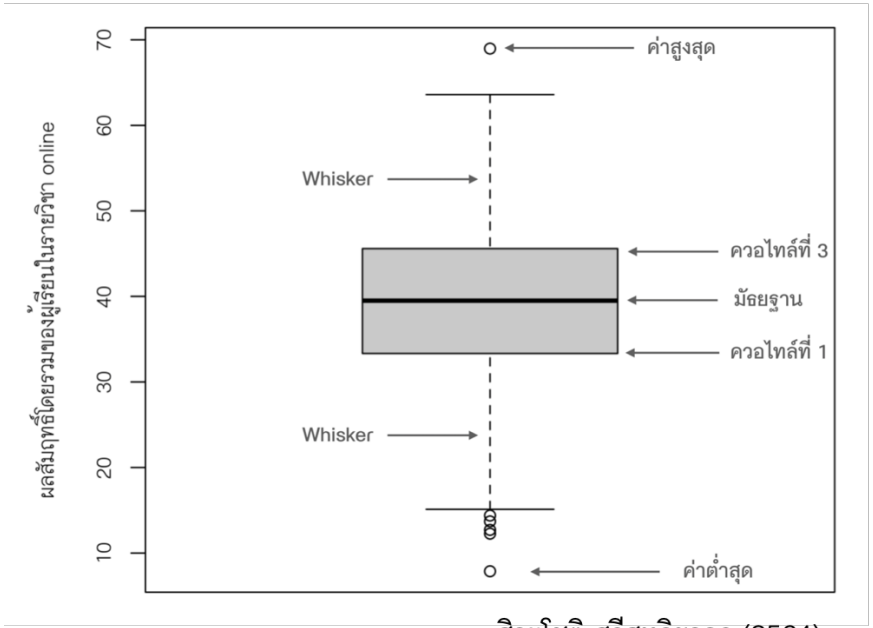


Pyramid Chart

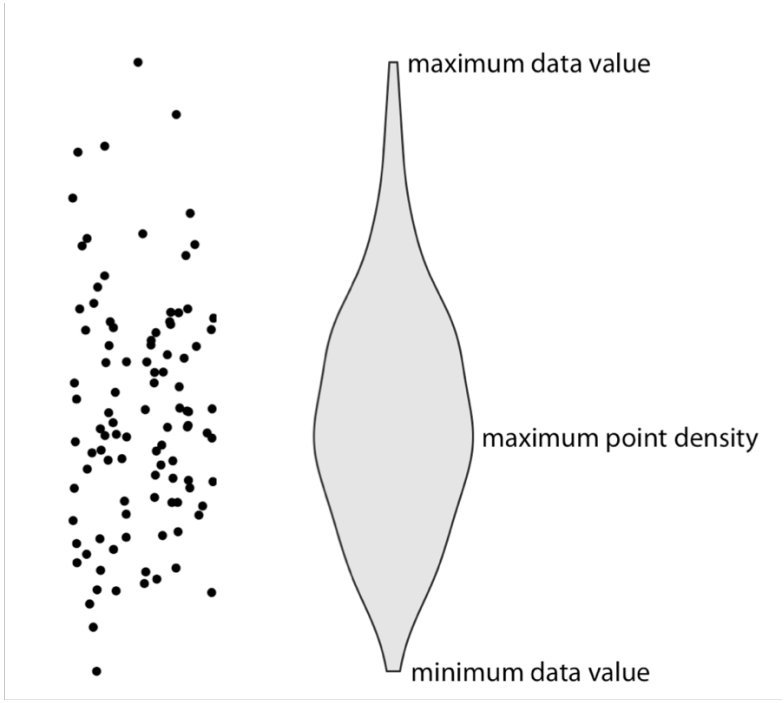
ใช้เปรียบเทียบการแจกแจงระหว่างสองกลุ่ม โดยสร้าง histogram สองตัวชนกัน



Boxplot, Jitter and Violin plot



สิวะโชติ ศรีสุทธิยากร (2564)

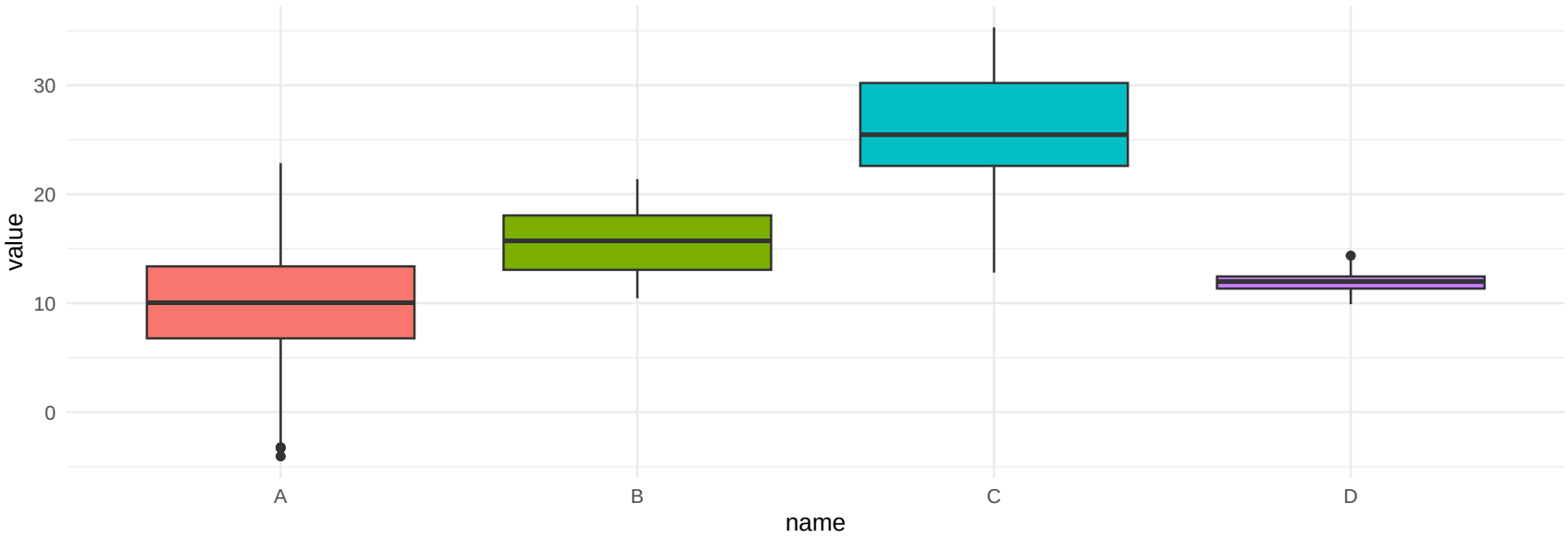


ข้อจำกัดของ boxplot คือสร้างจากค่าสถิติเพียง 5 ค่า ทำให้สูญเสีย information ไปพอสมควร ...

`geom_boxplot()`, `geom_jitter()`, and `geom_violin()`

Boxplot ของ GdpPercap จำแนกตามทวีป

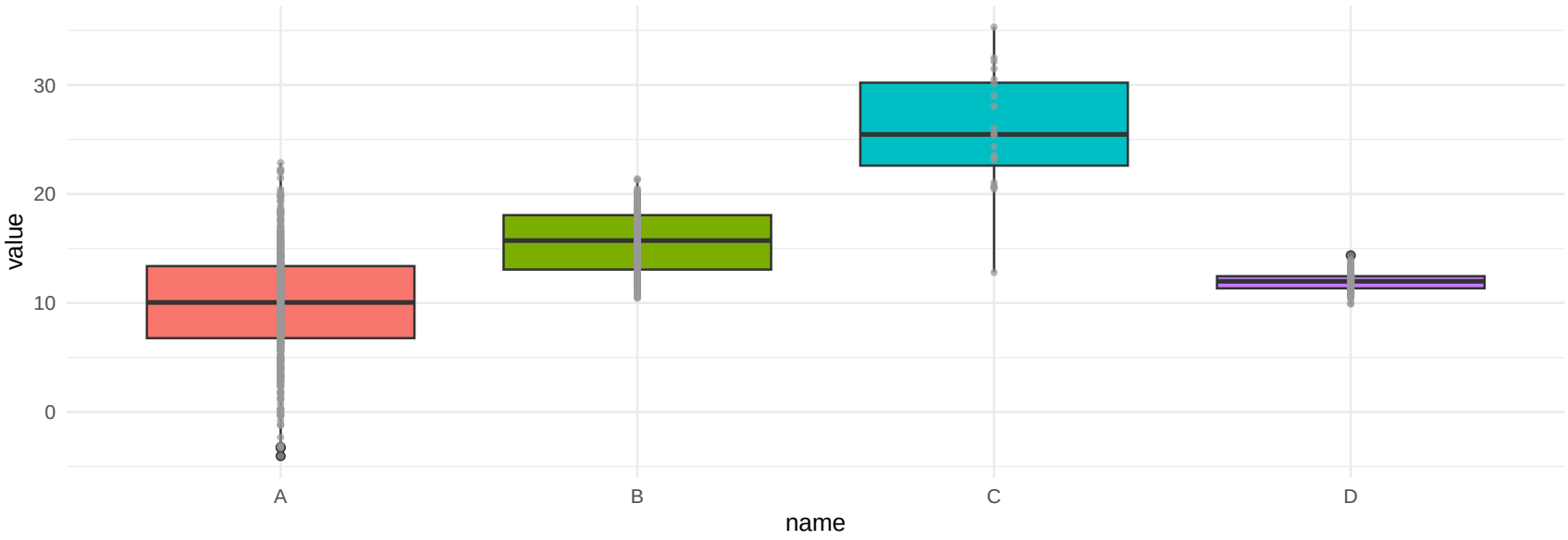
```
dat <- read.csv("boxplot.csv")
dat %>% ggplot(aes(x = name, y = value, fill = name))+
  geom_boxplot(show.legend = F)+
  theme_minimal()
```



Adding Points

Better ? Good visualization ?

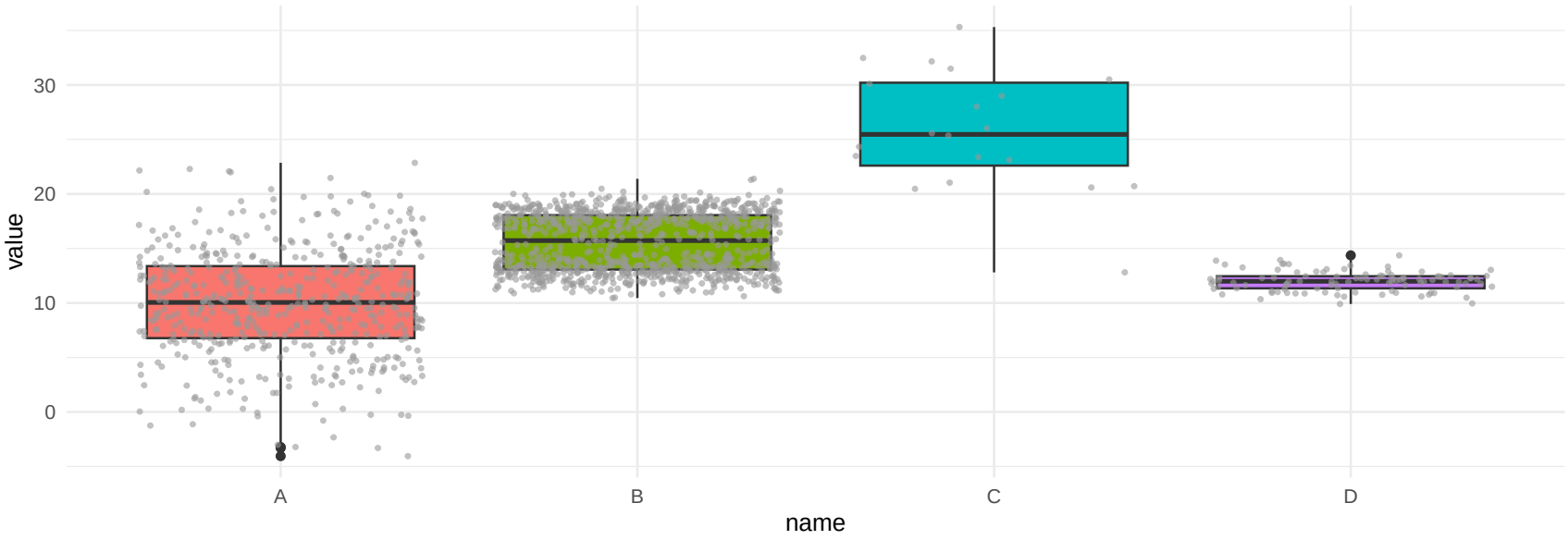
```
dat %>% ggplot(aes(x = name, y = value, fill = name))+
  geom_boxplot(show.legend = F)+
  geom_point(col = "grey60", alpha=0.7,
             size = 0.9, show.legend = F)+
  theme_minimal()
```



Adding Jitter

Better ? Good visualization ?

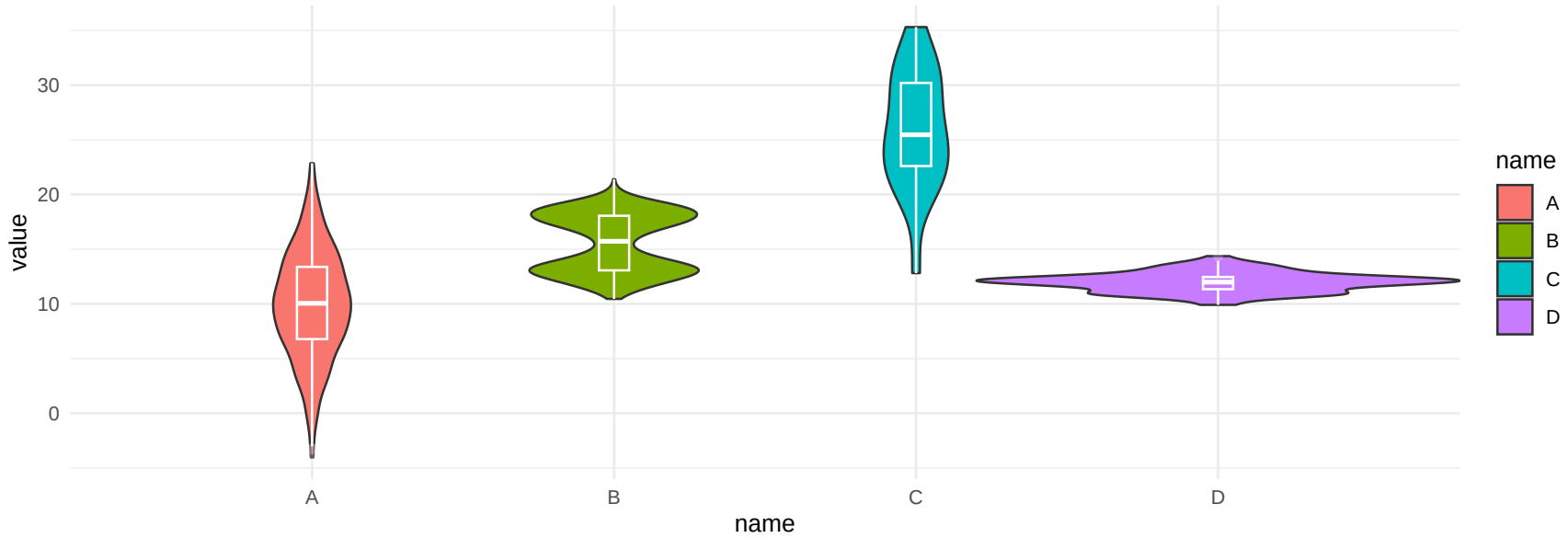
```
dat %>% ggplot(aes(x = name, y = value, fill = name))+
  geom_boxplot(show.legend = F)+
  geom_jitter(col = "grey60", alpha=0.6,
              size = 0.7, show.legend = F)+
  theme_minimal()
```



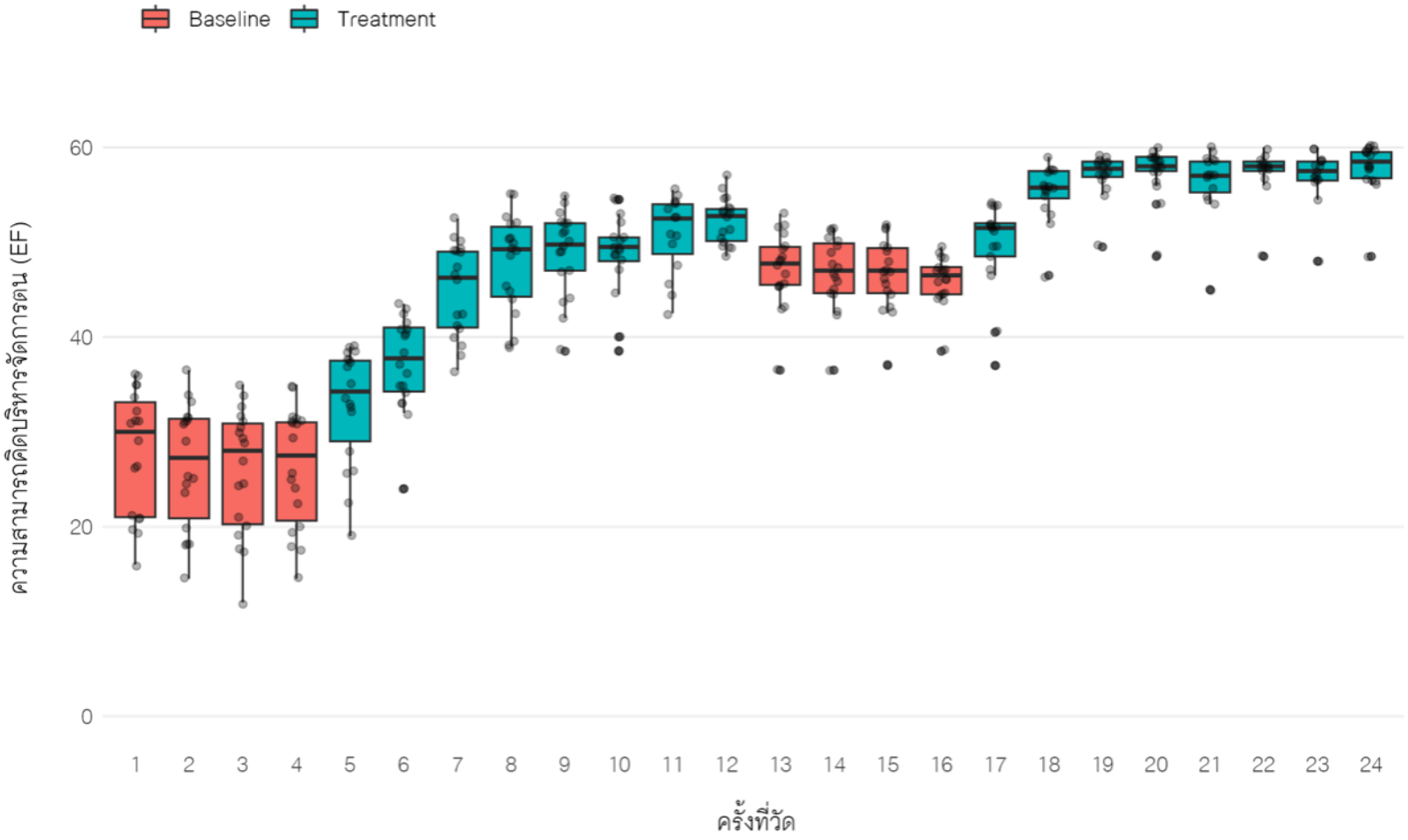
Using Violin plot

Better ? Good visualization ?

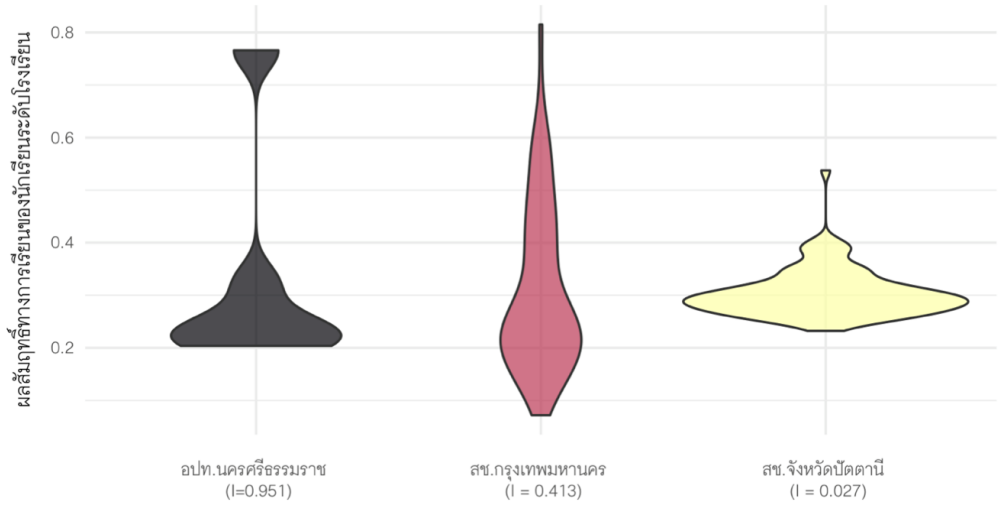
```
dat %>% ggplot(aes(x = name, y = value, fill = name))+
    geom_violin(width = 1.6)+
    geom_boxplot(color = "white", show.legend = F,
                 width = 0.1, alpha = 0.2)+
  theme_minimal()
```



Usecase: boxplot



Usecase: Violin plot



ภาพ 1 การแจกแจงของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนภายในสังกัดที่มีความเหลื่อมล้ำทางการศึกษาอยู่ในระดับสูงมาก ปานกลาง และต่ำ

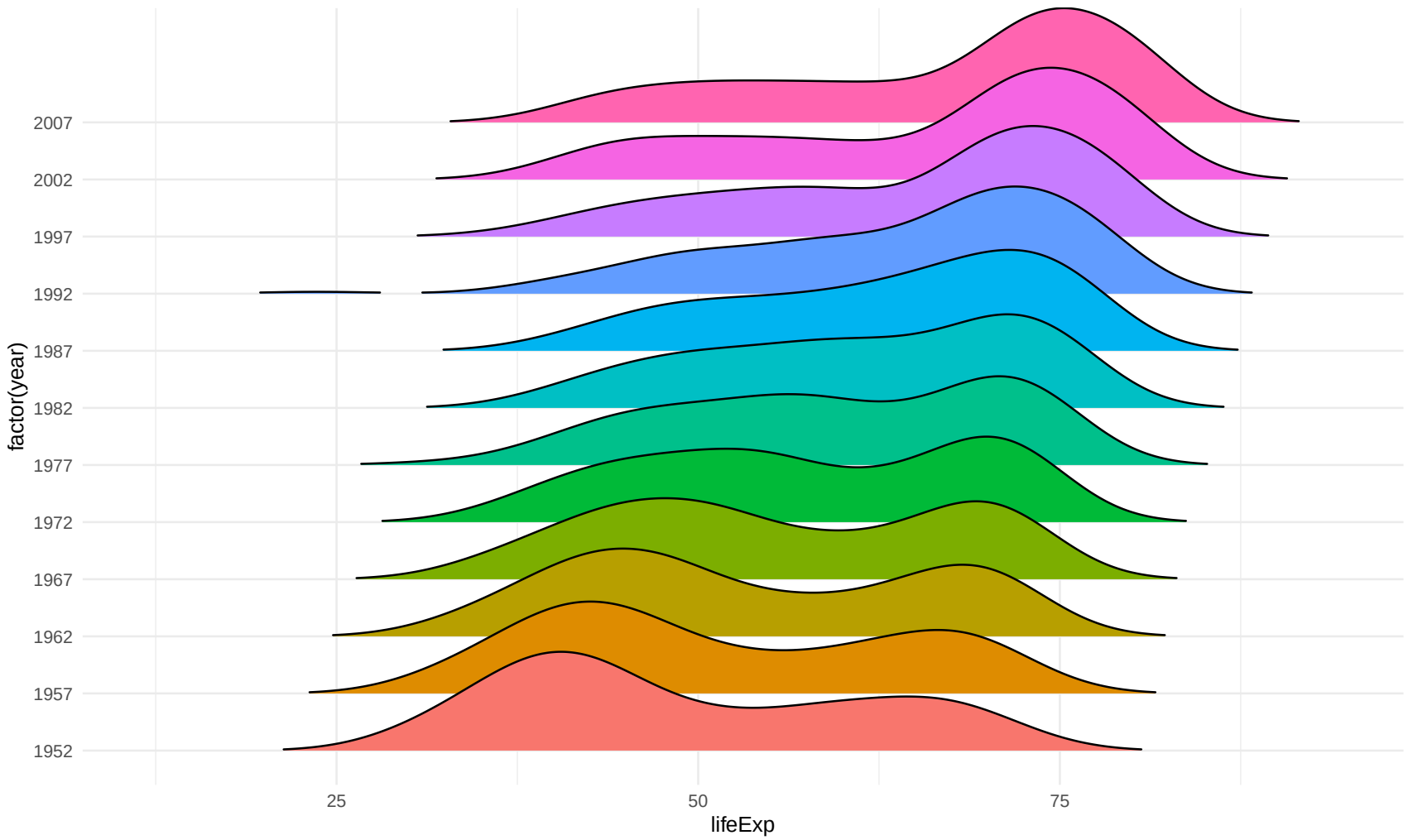
สิวะโชติ ศรีสุทธิยากร (2562)

Ridge line plot

เหมาะสำหรับเปรียบเทียบการแจกแจงของตัวแปรที่มีการเปลี่ยนแปลงตามเวลา แต่ก็สามารถใช้เปรียบเทียบระหว่างกลุ่มหลายกลุ่มได้ package-ggridges เป็น ggplot2 extension ตัวหนึ่งที่สามารถใช้สร้างแผนภาพดังกล่าวได้

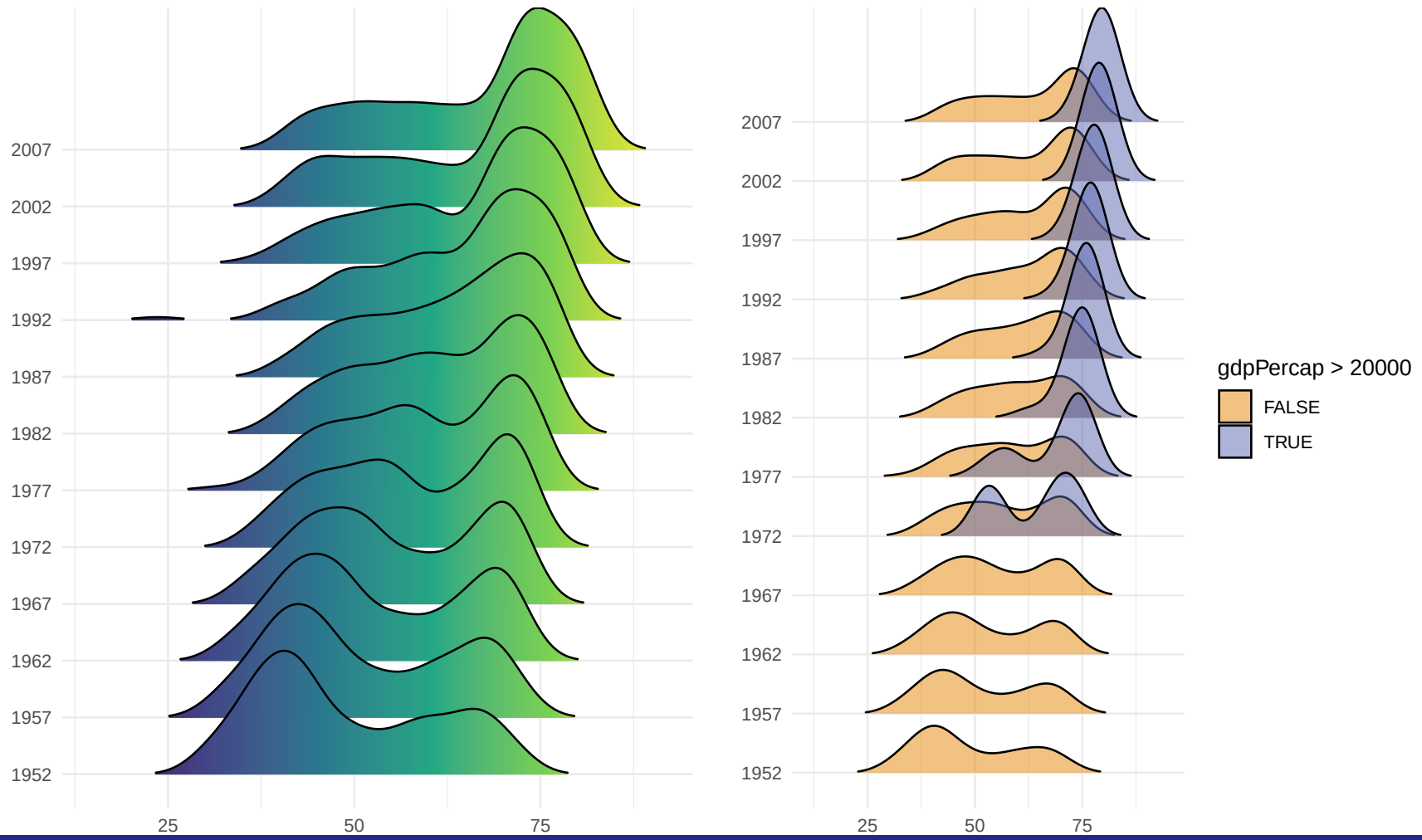
- `geom_ridgeline()`
- `geom_ridgeline_gradient()`
- `geom_density_ridges()`
- `geom_density_ridges_gradient()`

Ridge line plot



Ridge line plot

```
geom_density_ridges_gradient()
```

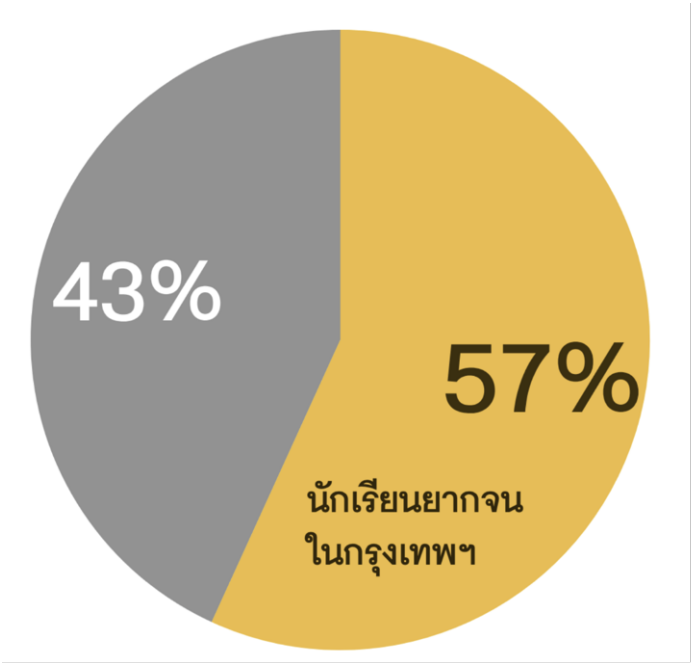


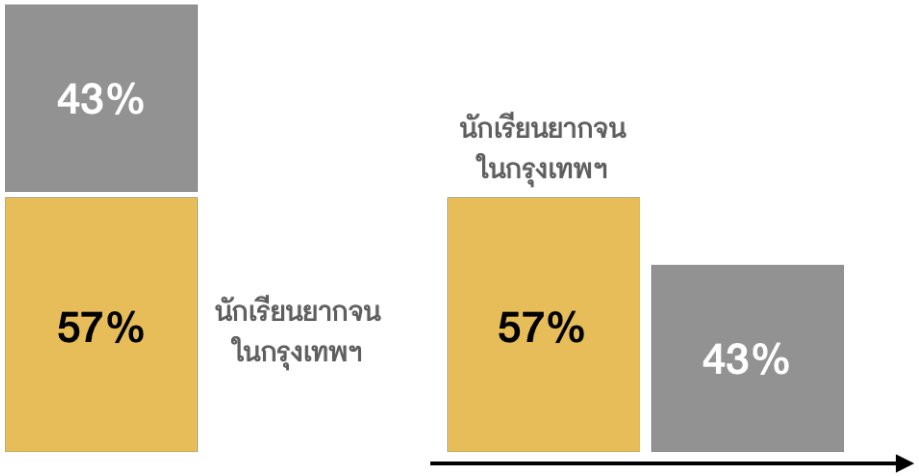
Visualizing Proportions

Visualizing Proportions

ที่มา : Wike (2019)

Pie Charts vs Bar Charts





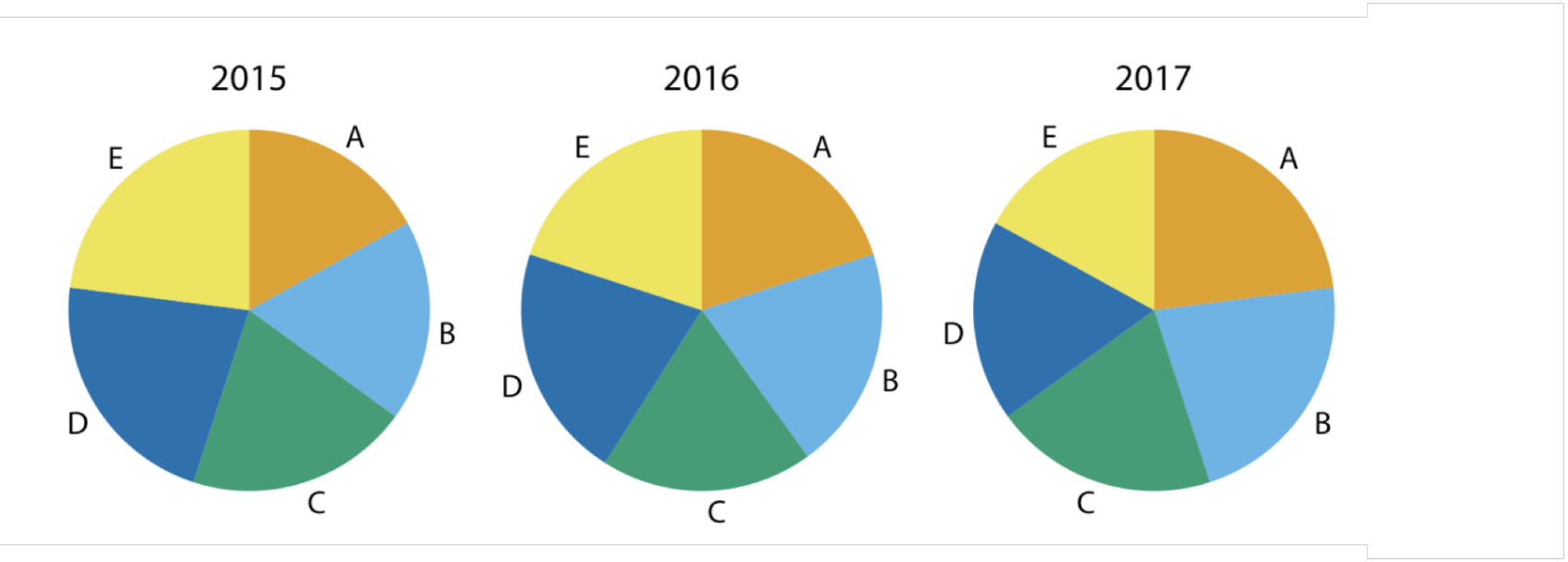
ที่มา : สิวะโชติ ศรีสุทธิยากร (2565)

Pie Charts vs Bar Charts

วัตถุประสงค์	Pie Chart	100% Stacked Bar	Side Bars
เปรียบเทียบส่วนย่อยที่สนใจกับส่วนทั้งหมด	Yes	Yes	No
นำเสนอส่วนย่อยในรูปสัดส่วนอย่างง่าย เช่น 1/2, 1/3, 1/4	Yes	No	No
เปรียบเทียบปริมาณระหว่างส่วนย่อยหลาย ๆ ส่วน	No	No	Yes
เปรียบเทียบสัดส่วน หลายชุดตามเวลา	No	Yes	No

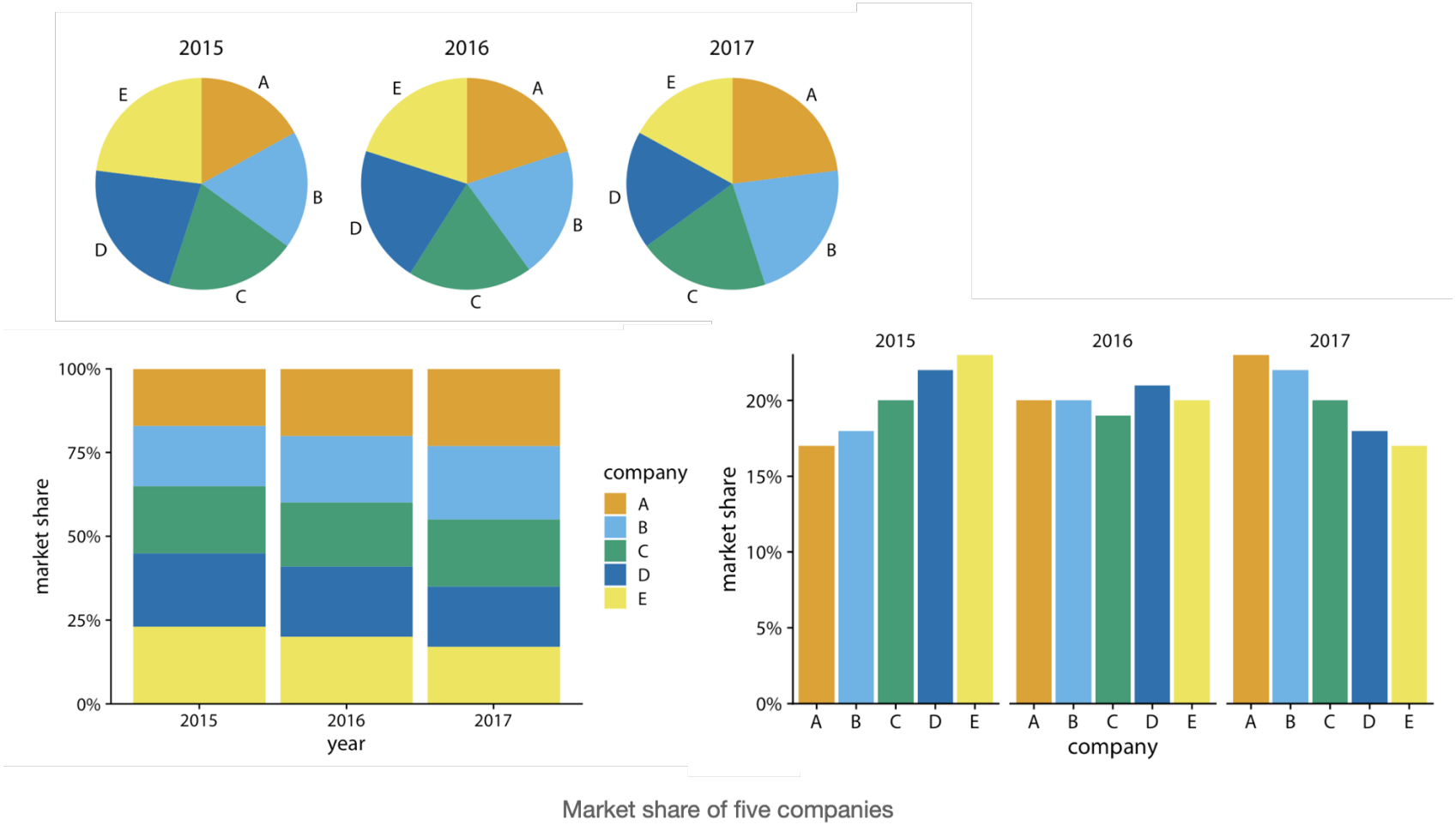
ข้อควรพิจารณา 1

วัตถุประสงค์ : ต้องการเปรียบเทียบยอดขายสินค้า 5 ชนิด ระหว่างปี 2015 - 2017



ที่มา : Wike (2019)

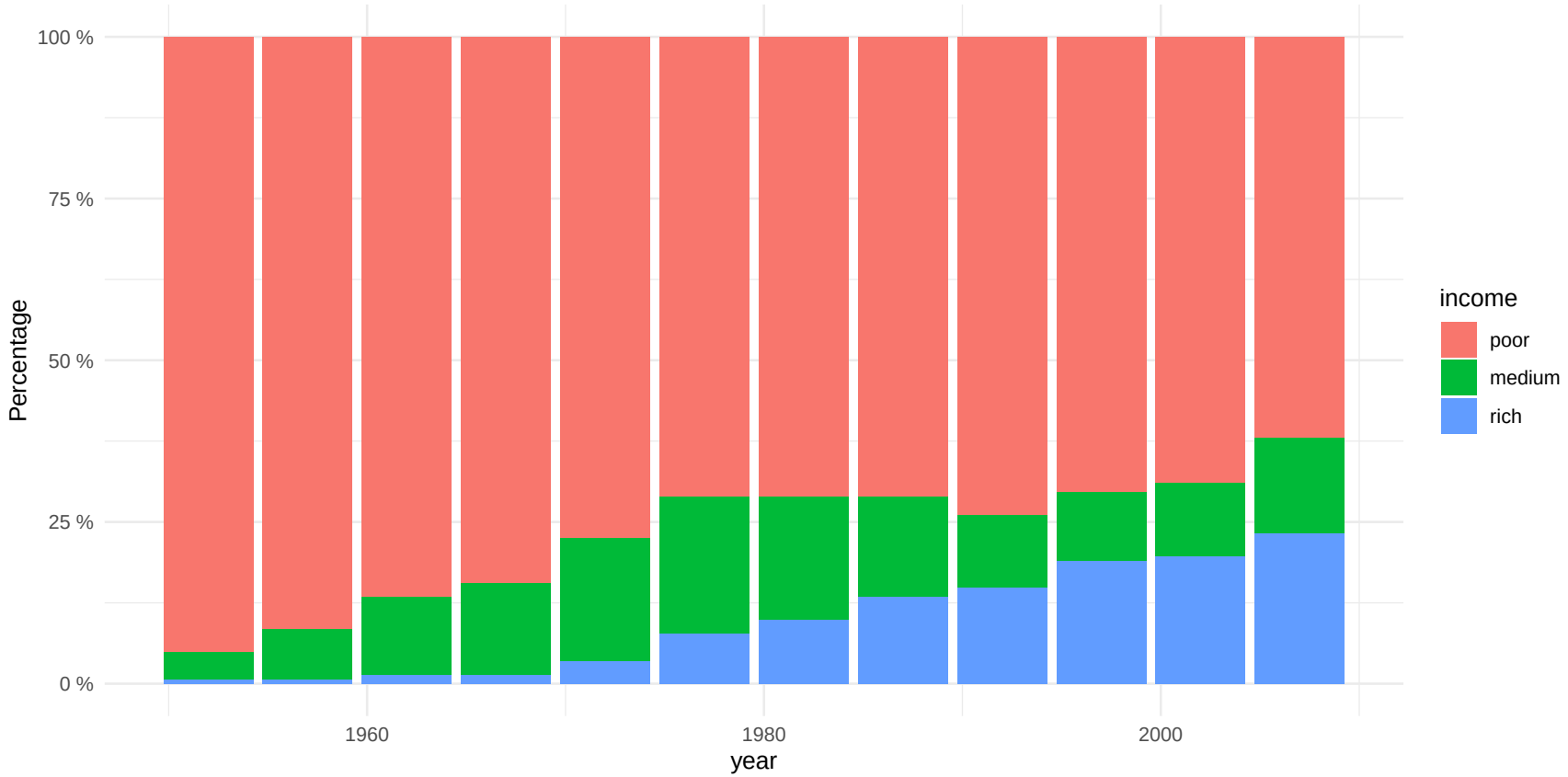
ข้อควรพิจารณา 1



ที่มา : Wike (2019)

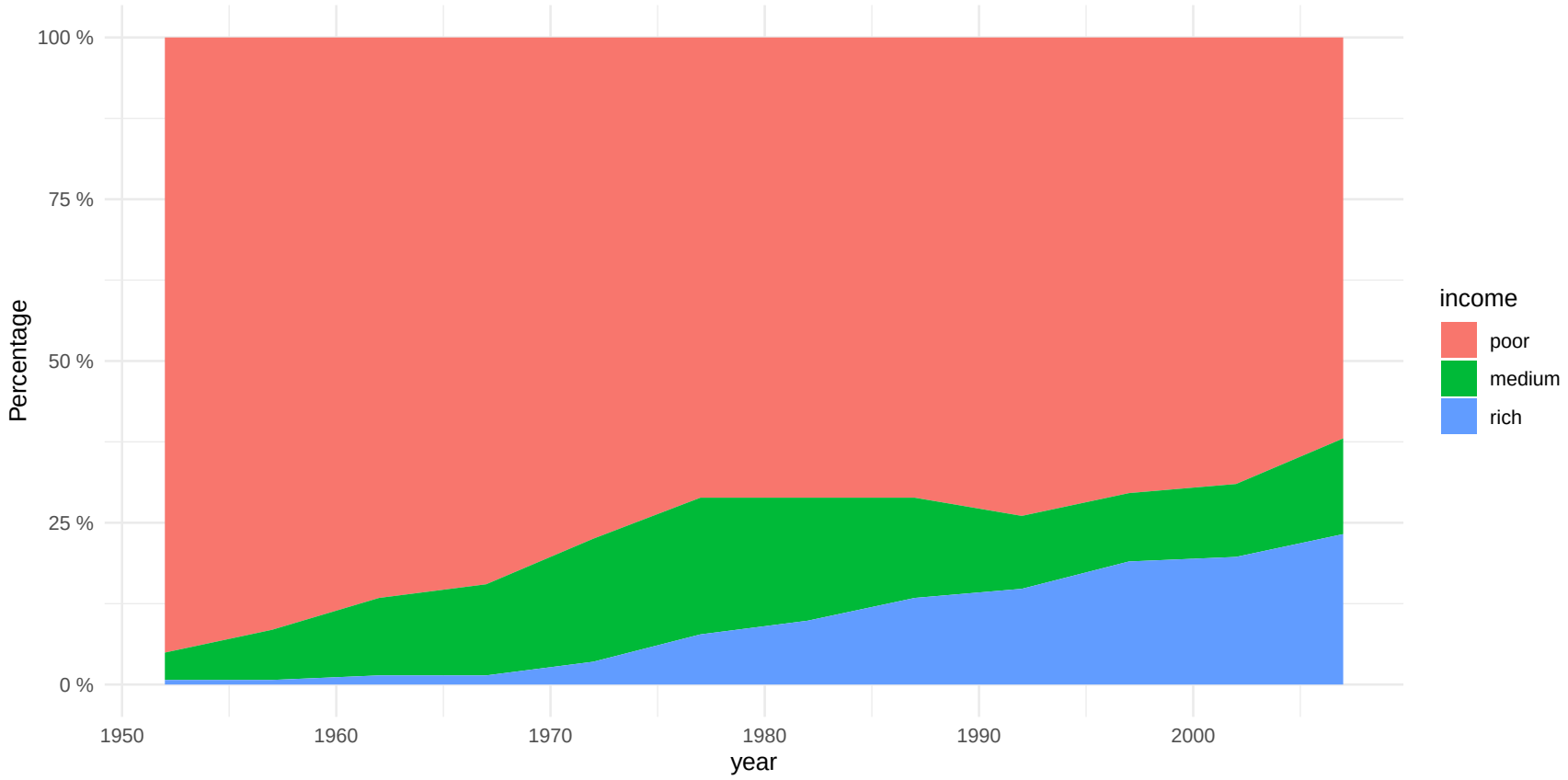
100% Stacked barplot (again)

100% stacked barplot ยังสามารถใช้วิเคราะห์แนวโน้มตามเวลาได้ด้วย



100% Stacked areaplot

เนื่องจากเวลาเป็นค่าต่อเนื่อง เราสามารถเปลี่ยน geometric จากแท่งสี่เหลี่ยมเป็นพื้นที่ได้ โดยใช้ `geom_area()` ดังรูป

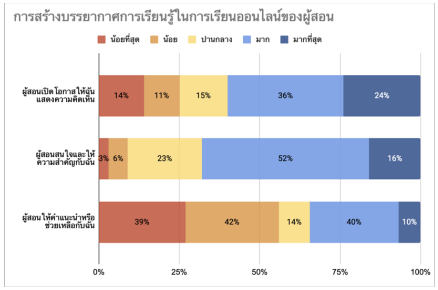


100% stacked barplot: Likert plot

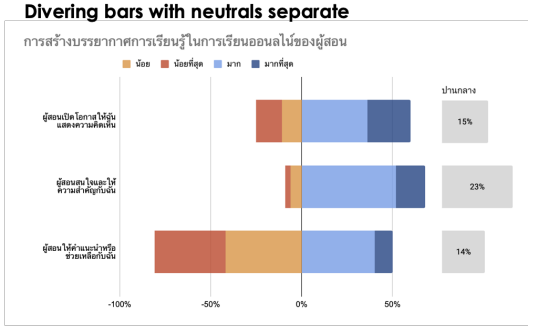
Stacked bar plot สามารถมาประยุกต์ใช้กับการนำเสนอ
สถิติพื้นฐานของข้อคำถามแบบ rating scale ได้ ดังตัวอย่างต่อไปนี้

item	น้อยที่สุด	น้อย	ปานกลาง	มาก	มากที่สุด
ผู้สอนเปิดโอกาสให้ฉัน แสดงความคิดเห็น	14.00%	11.00%	15.00%	36.00%	24.00%
ผู้สอนสนใจและให้ ความสำคัญกับฉัน	3.00%	6.00%	23.00%	52.00%	16.00%
ผู้สอนให้คำแนะนำหรือ ช่วยเหลือกับฉัน	39.00%	42.00%	14.00%	40.00%	10.00%

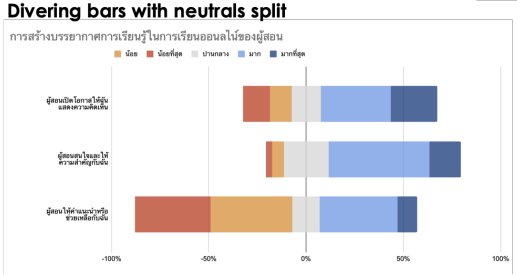
100% stacked bar



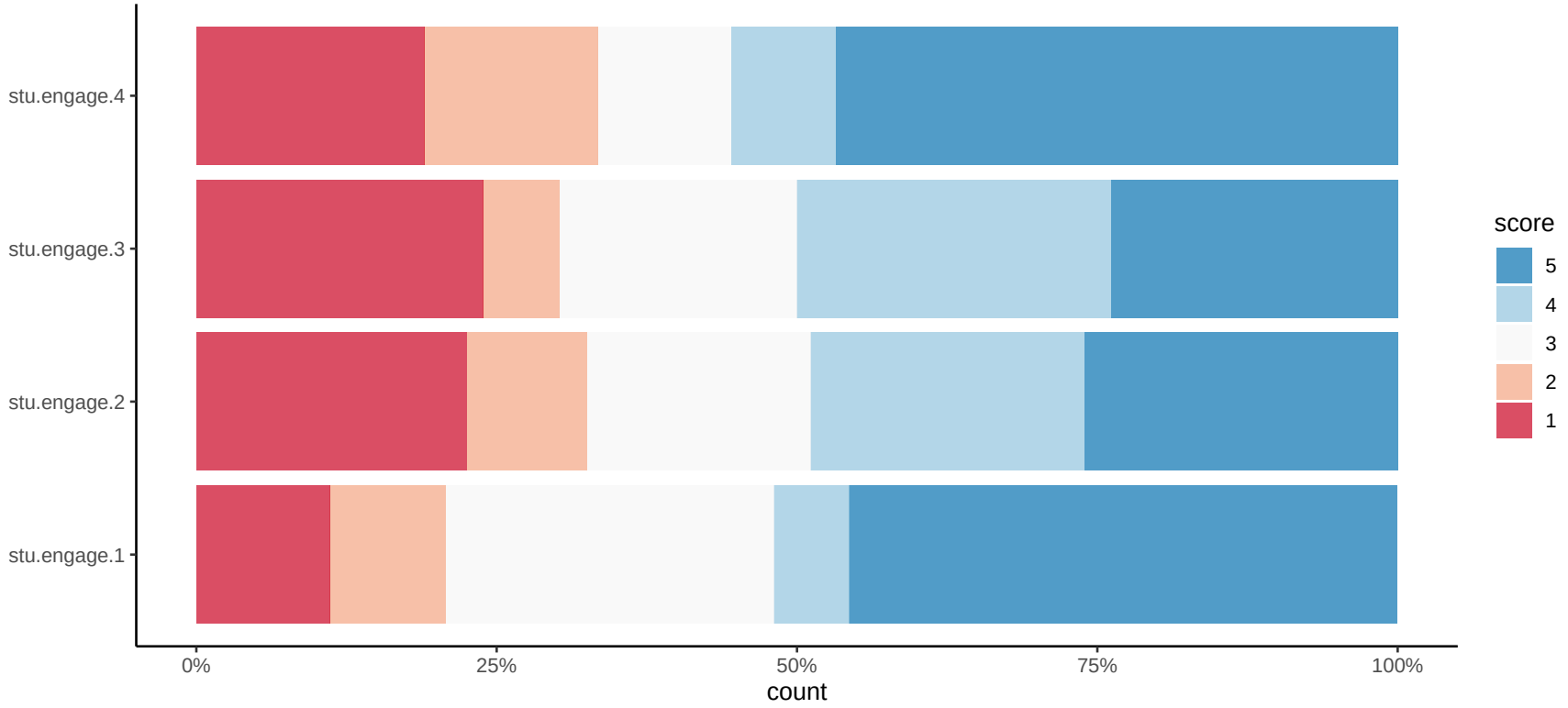
Divering bars with neutrals separate



Divering bars with neutrals split



Likert plot


stu.engage.4
stu.engage.3
stu.engage.2
stu.engage.1
0%
25%
50%
75%
100%
count
score
5
4
3
2
1

Visualizing Nested Proportion

Visualizing Nested Proportion

ใช้ในกรณีที่มีการแจกแจงความถี่ของตัวแปรจัดประเภทหลายทาง ดังตัวอย่างตารางด้านล่าง

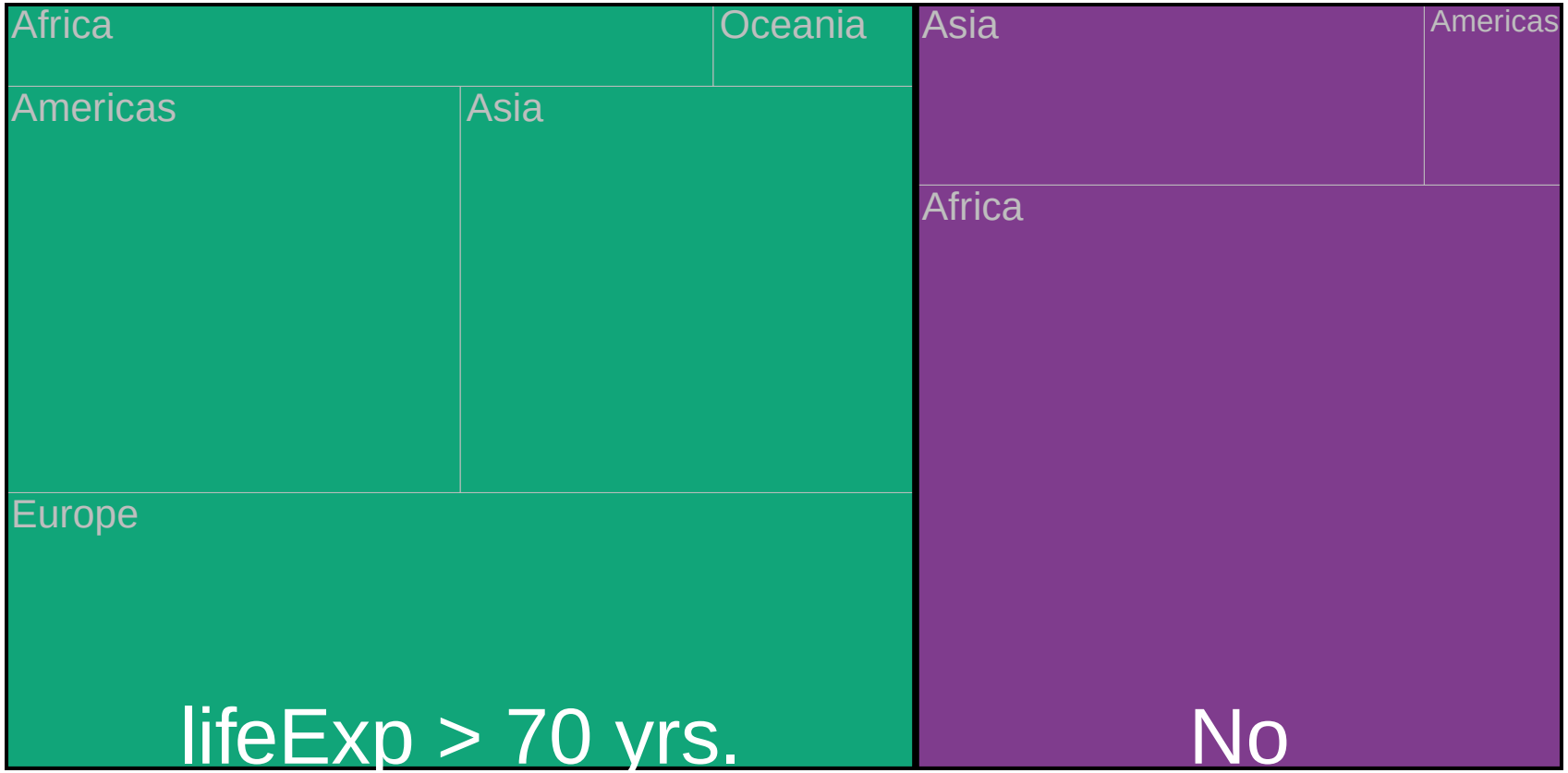
```
# A tibble: 8 x 3
# Groups:   continent, lifeExp > 70 [8]
  continent `lifeExp > 70`     n
  <fct>     <lgl>          <int>
1 Africa    FALSE             45
2 Africa    TRUE               7
3 Americas  FALSE              3
4 Americas  TRUE              22
5 Asia      FALSE             11
6 Asia      TRUE              22
7 Europe    TRUE              30
8 Oceania   TRUE               2
```

Visualizing Nested Proportion: Treemap

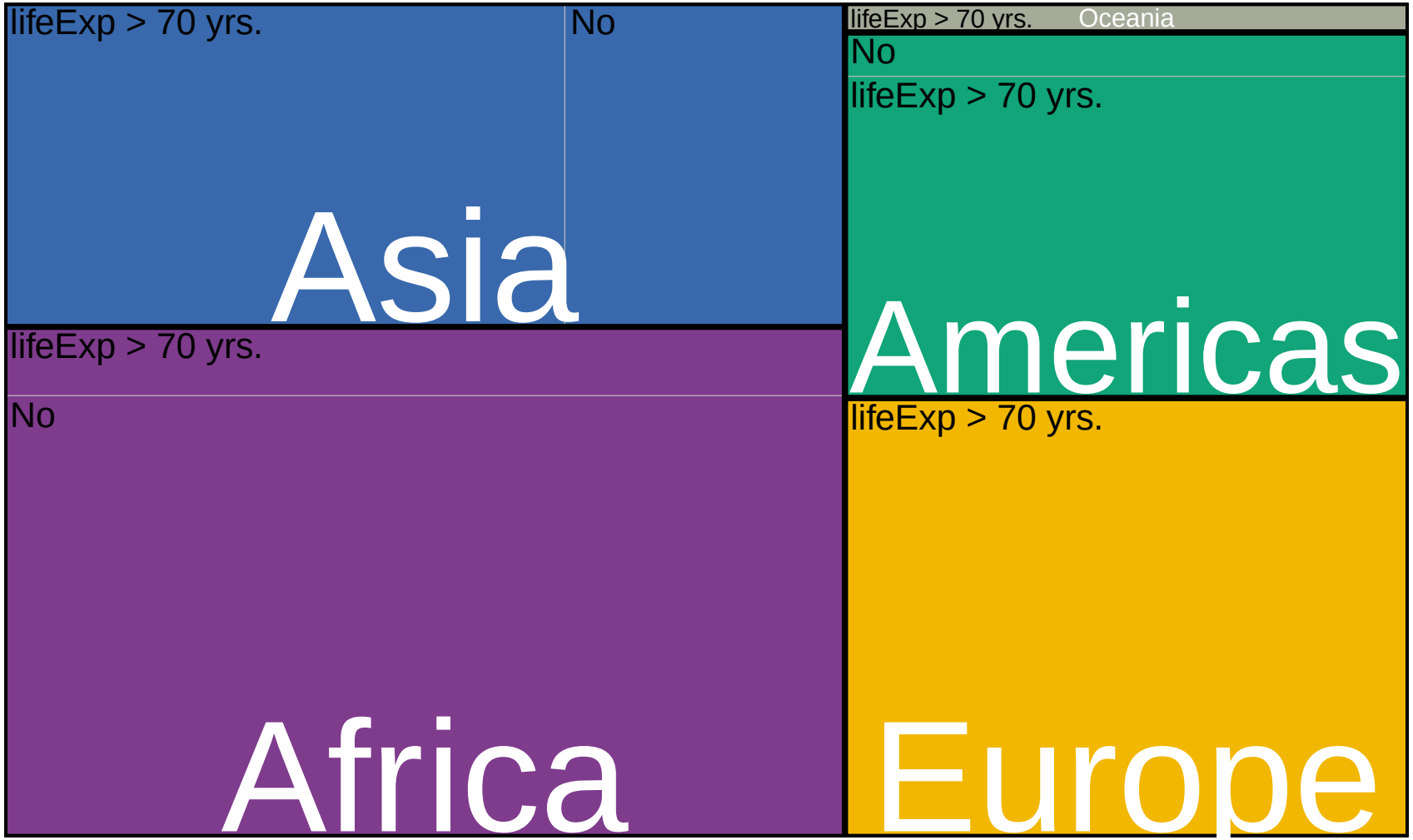
แผนภาพ Treemap เป็น visualization
ตัวหนึ่งที่มีประสิทธิภาพในการนำเสนอการแจกแจงความถี่หลายทางดังกล่าว

```
install.packages("devtools")
library(devtools)
install_github("wilkox/treemapify")
library(treemapify)
```

Visualizing Nested Proportion: Treemap



Visualizing Nested Proportion: Treemap



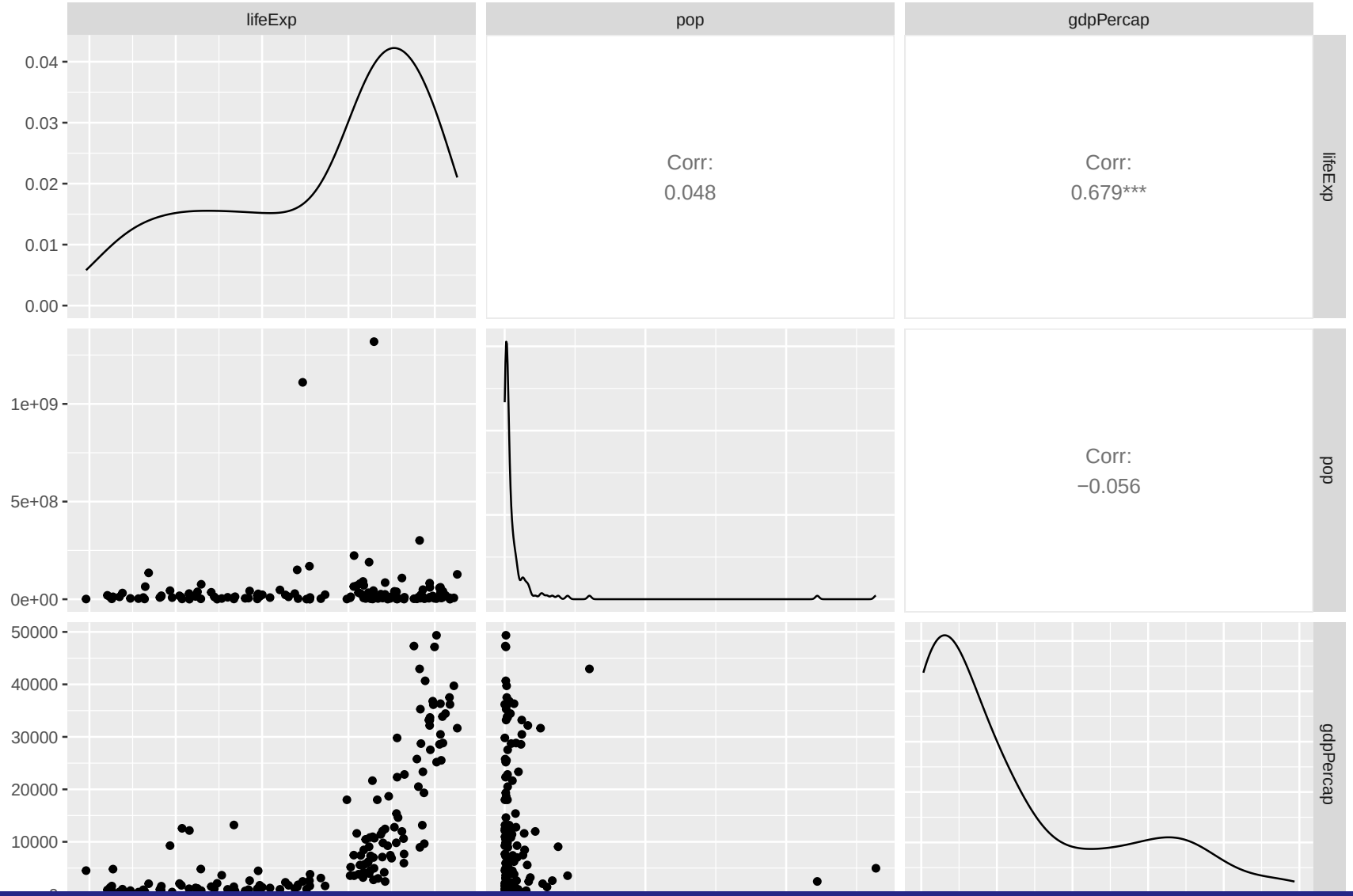
Visualizing Relations & Trends

Scatterplot matrix

การวิเคราะห์ความสัมพันธ์ด้วยทัศนภาพข้อมูล สามารถทำได้หลายวิธี
วิธีการหลักคือ scatter plot ที่ได้กล่าวไปแล้ว นอกจากนี้ยังมี scatter plot matrix ที่มีประโยชน์ในการสำรวจข้อมูล

```
library(GGally)
gapminder %>%
  filter(year == "2007") %>%
  ggpairs(columns=4:6)
```

Scatterplot matrix



Correlation plot

correlation matrix สามารถนำมาสร้างเป็น heatmap ในลักษณะต่าง ๆ ได้ ผู้วิเคราะห์สามารถใช้ `geom_tile()` ก็ได้ หรือจะใช้ ggextension ได้แก่ package `ggcorrplot` ซึ่งทำได้สะดวกกว่า ดังนี้

```
#install.packages("ggcorrplot")
library(ggcorrplot)
```

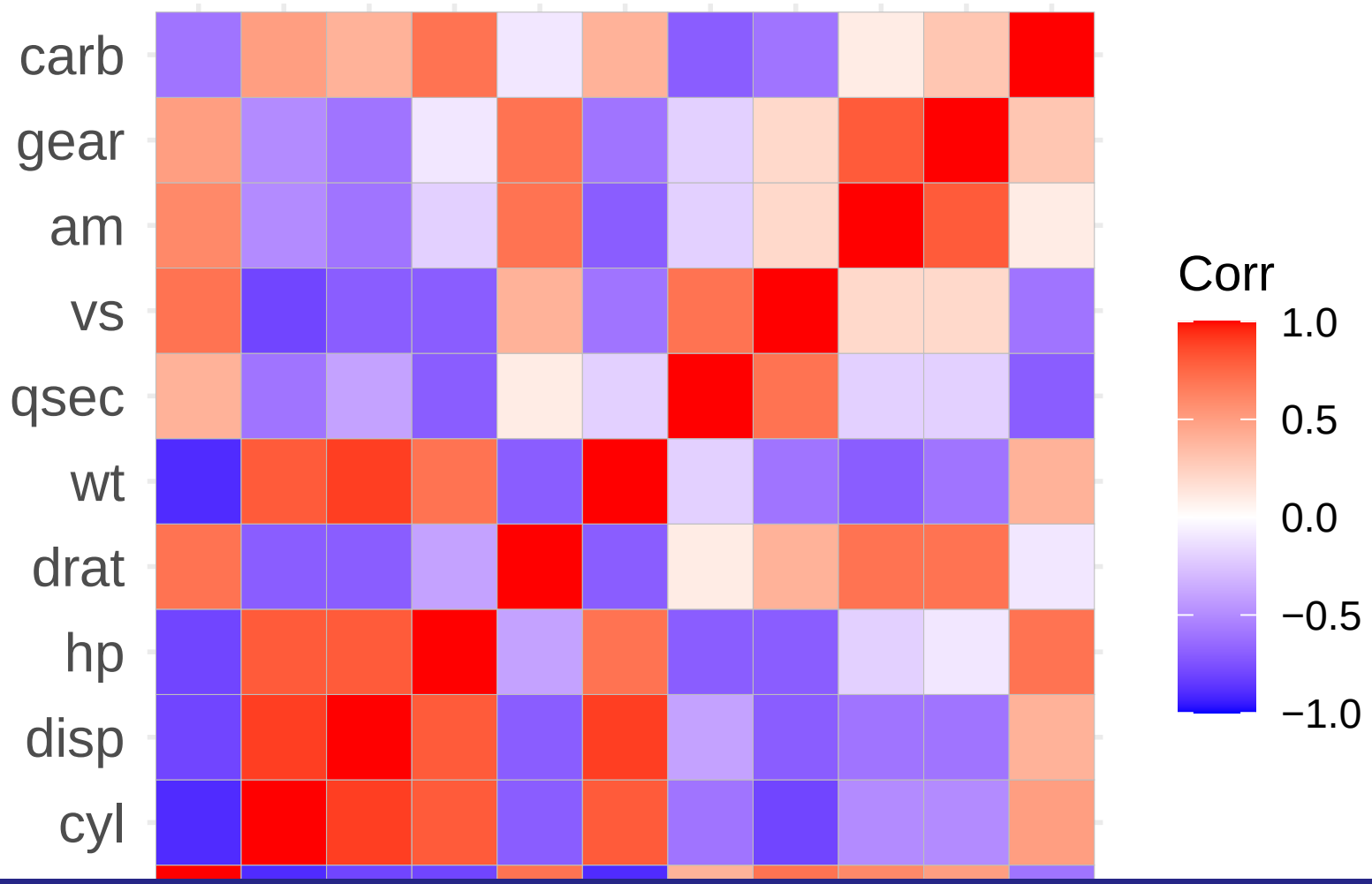
Correlation plot

```r
# Compute a correlation matrix
data(mtcars)
corr <- round(cor(mtcars), 1)
head(corr[, 1:6])
```

```
      mpg  cyl disp   hp drat   wt
mpg   1.0 -0.9 -0.8 -0.8  0.7 -0.9
cyl  -0.9  1.0  0.9  0.8 -0.7  0.8
disp -0.8  0.9  1.0  0.8 -0.7  0.9
hp   -0.8  0.8  0.8  1.0 -0.4  0.7
drat  0.7 -0.7 -0.7 -0.4  1.0 -0.7
wt   -0.9  0.8  0.9  0.7 -0.7  1.0
```

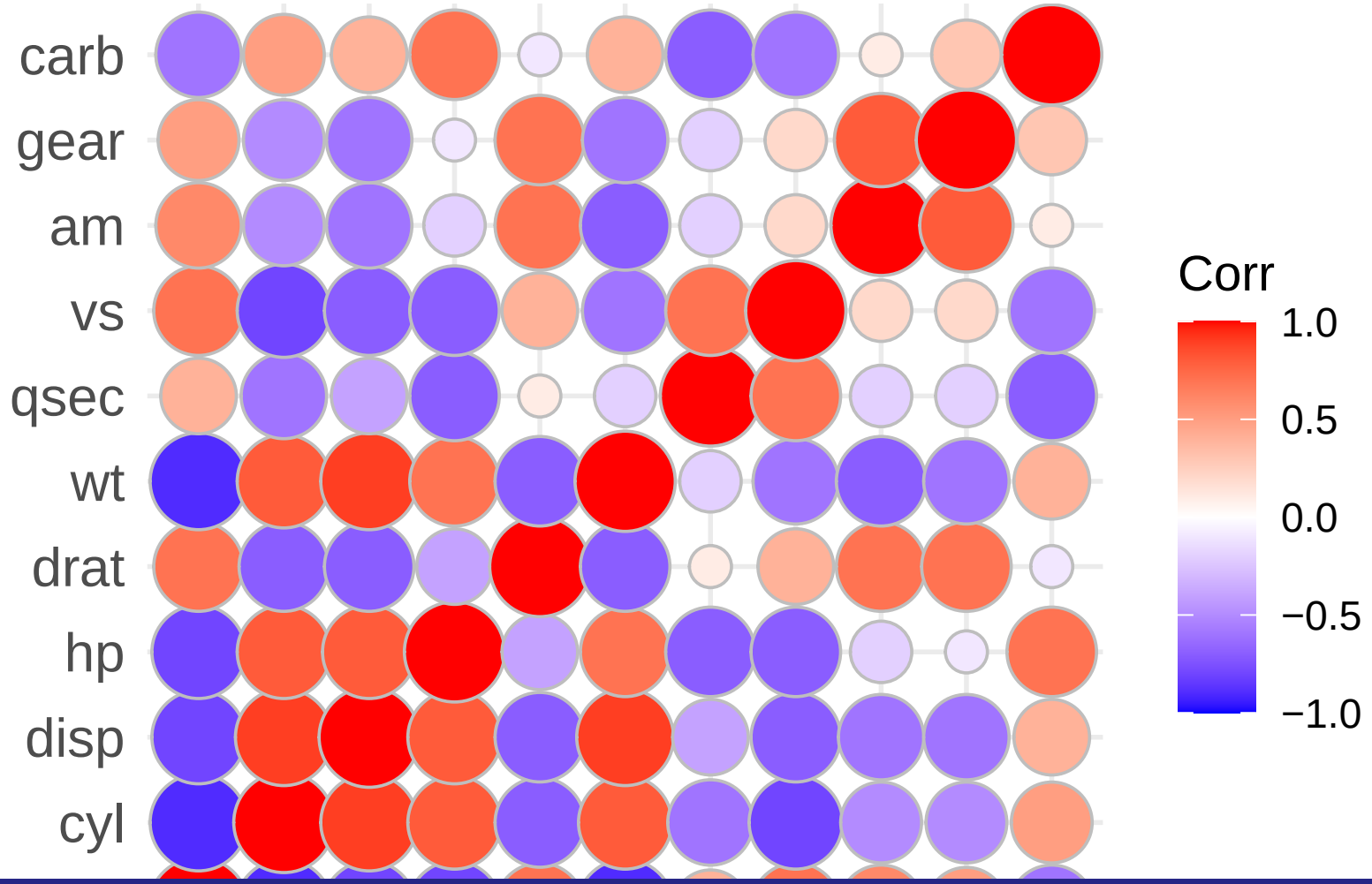
Correlation plot

```
ggcorrplot(corr)
```



Correlation plot

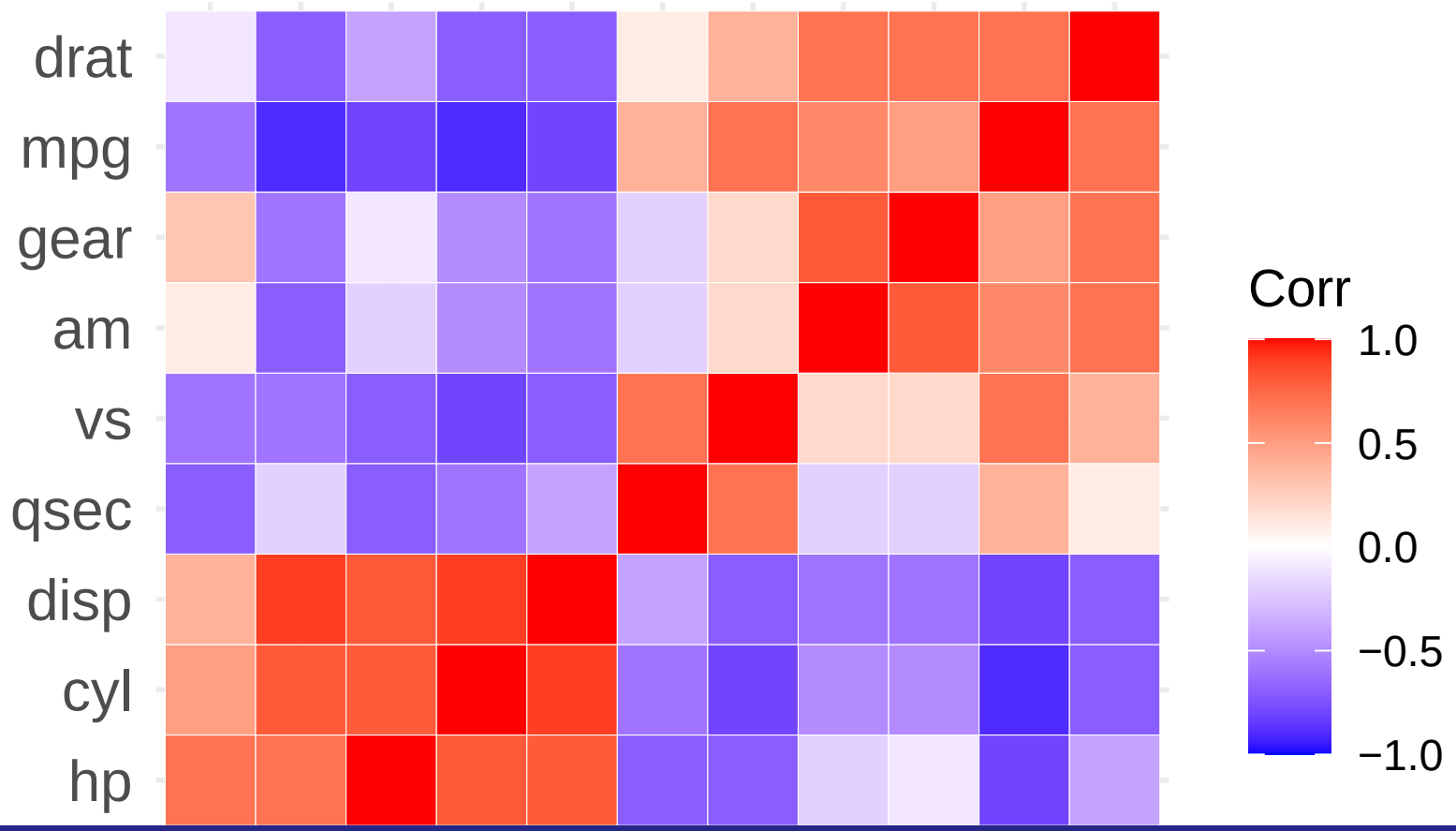
```
ggcorrplot(corr, method = "circle")
```



Correlation plot

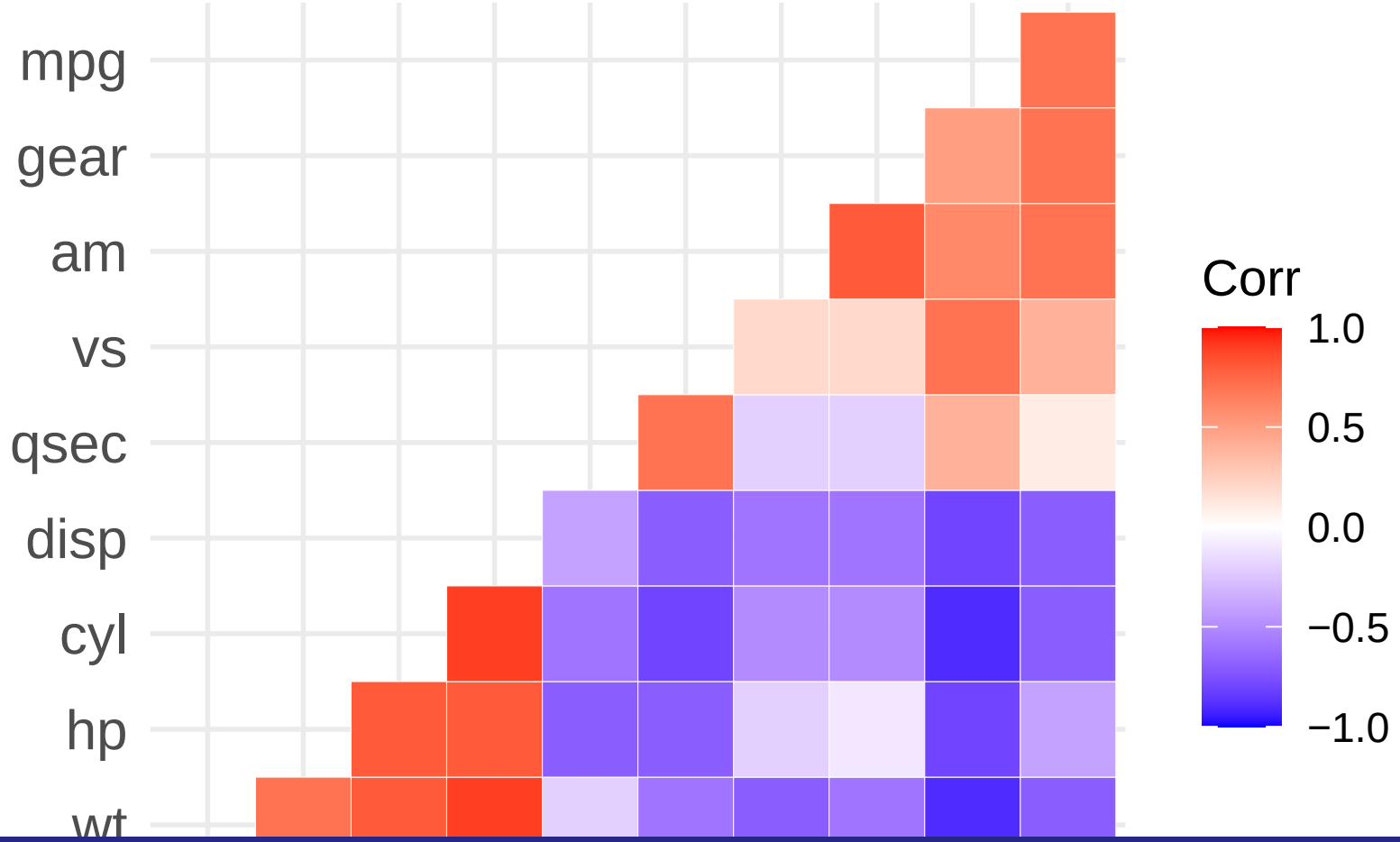
hierarchical clustering

```
ggcorrplot(corr, hc.order = TRUE, outline.col = "white")
```



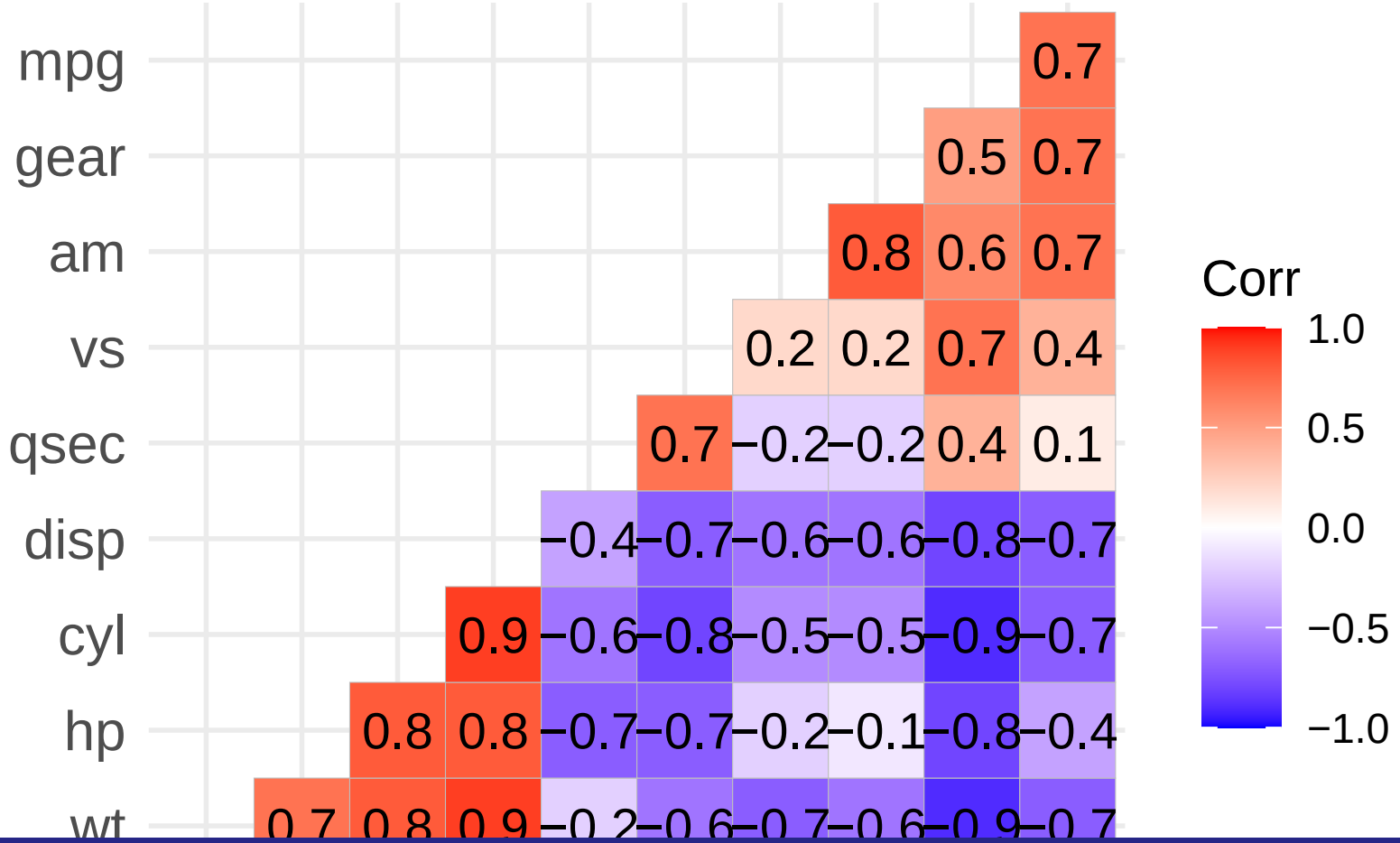
Correlation plot

```r
ggcorrplot(corr, hc.order = TRUE, type = "lower",
     outline.col = "white")
```



Correlation plot

```
ggcorrplot(corr, hc.order = TRUE, type = "lower",
    lab = TRUE)
```



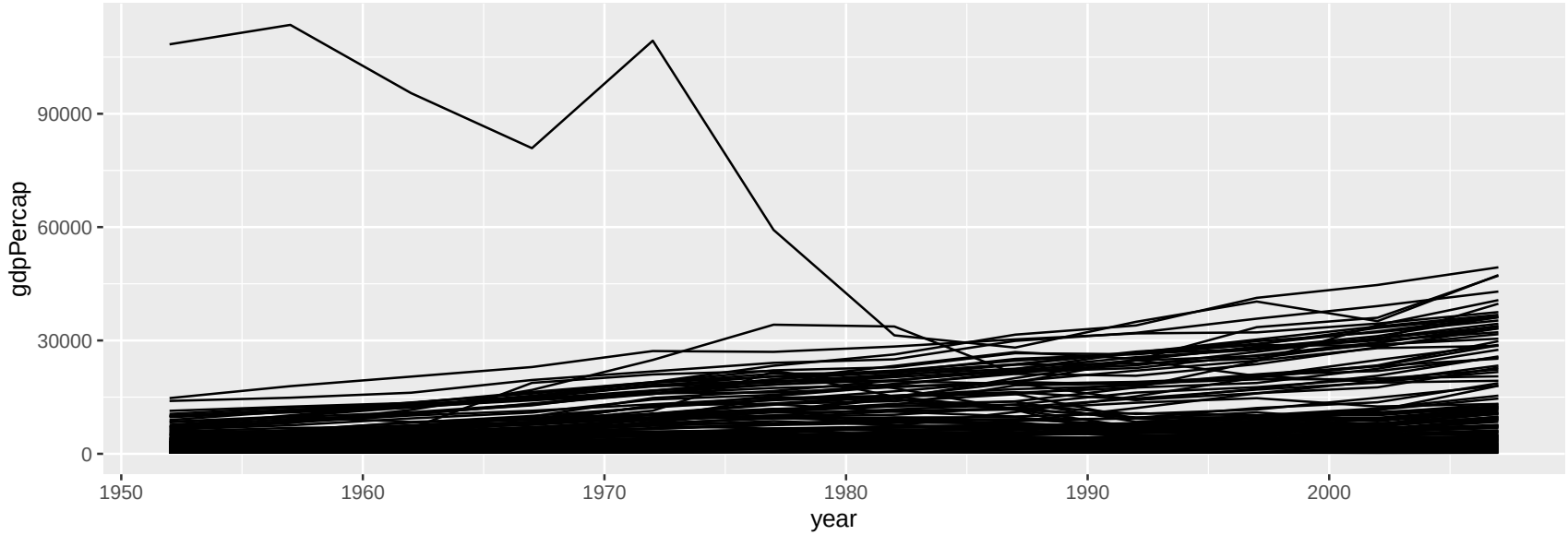
Trends

- line graph
- dot plot
- slope plot

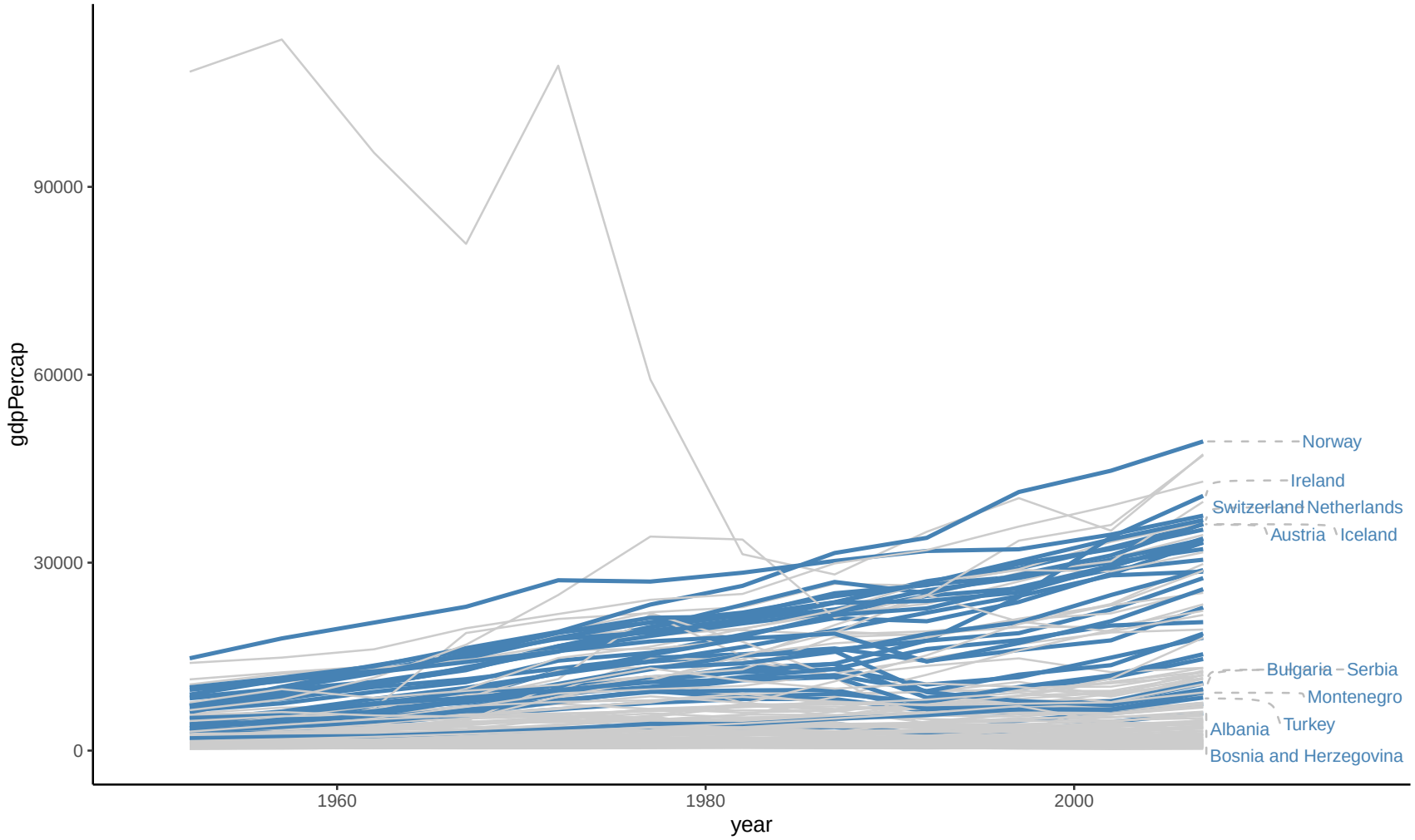
Line graph

วิเคราะห์แนวโน้มของ gdpPercap ของประเทศต่าง ๆ ตั้งแต่ปี 1952 ถึง 2007

```
gapminder %>%
  ggplot(aes(x = year,
             y=gdpPercap,
             group = country))+
  geom_line()
```

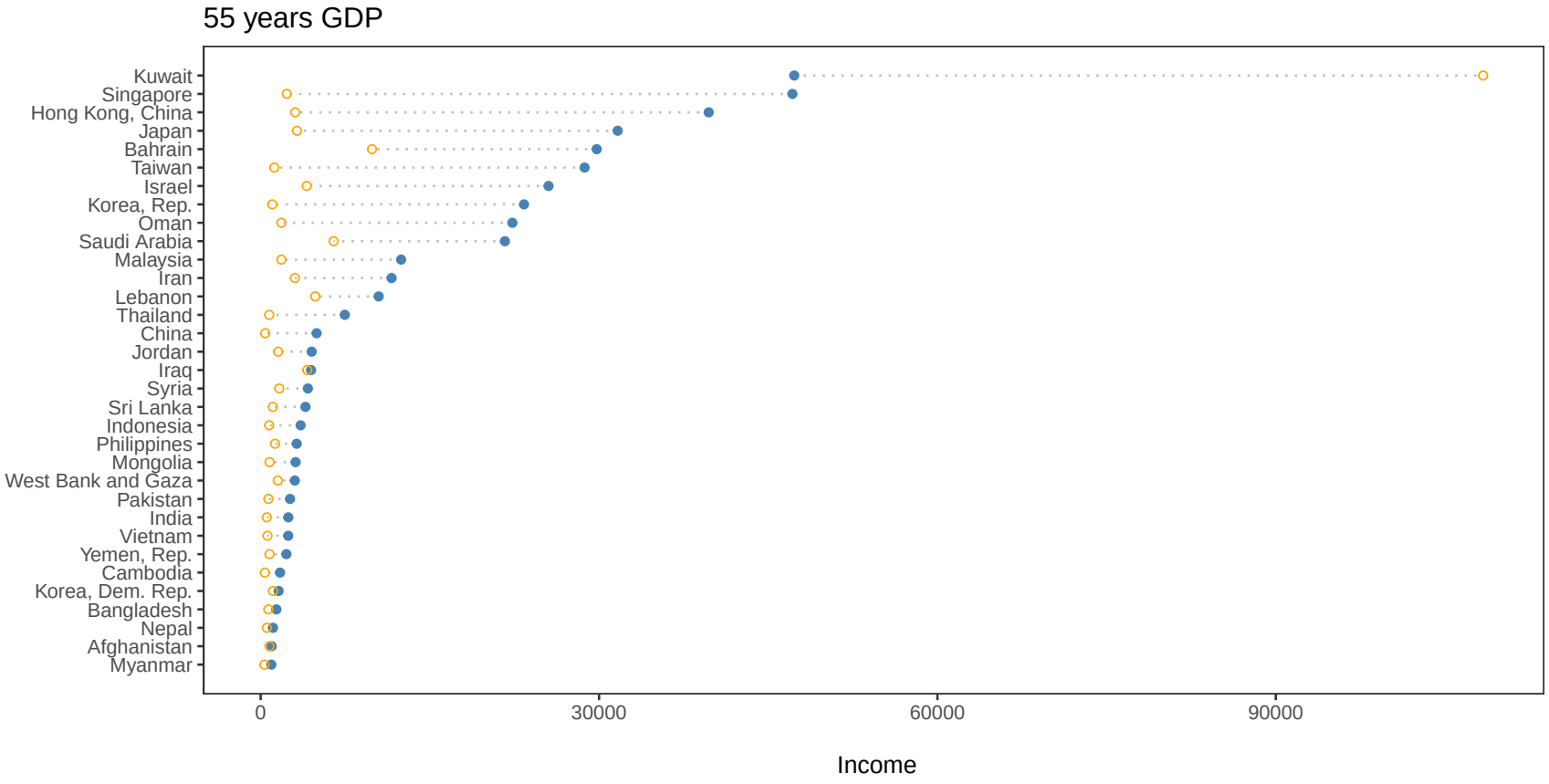
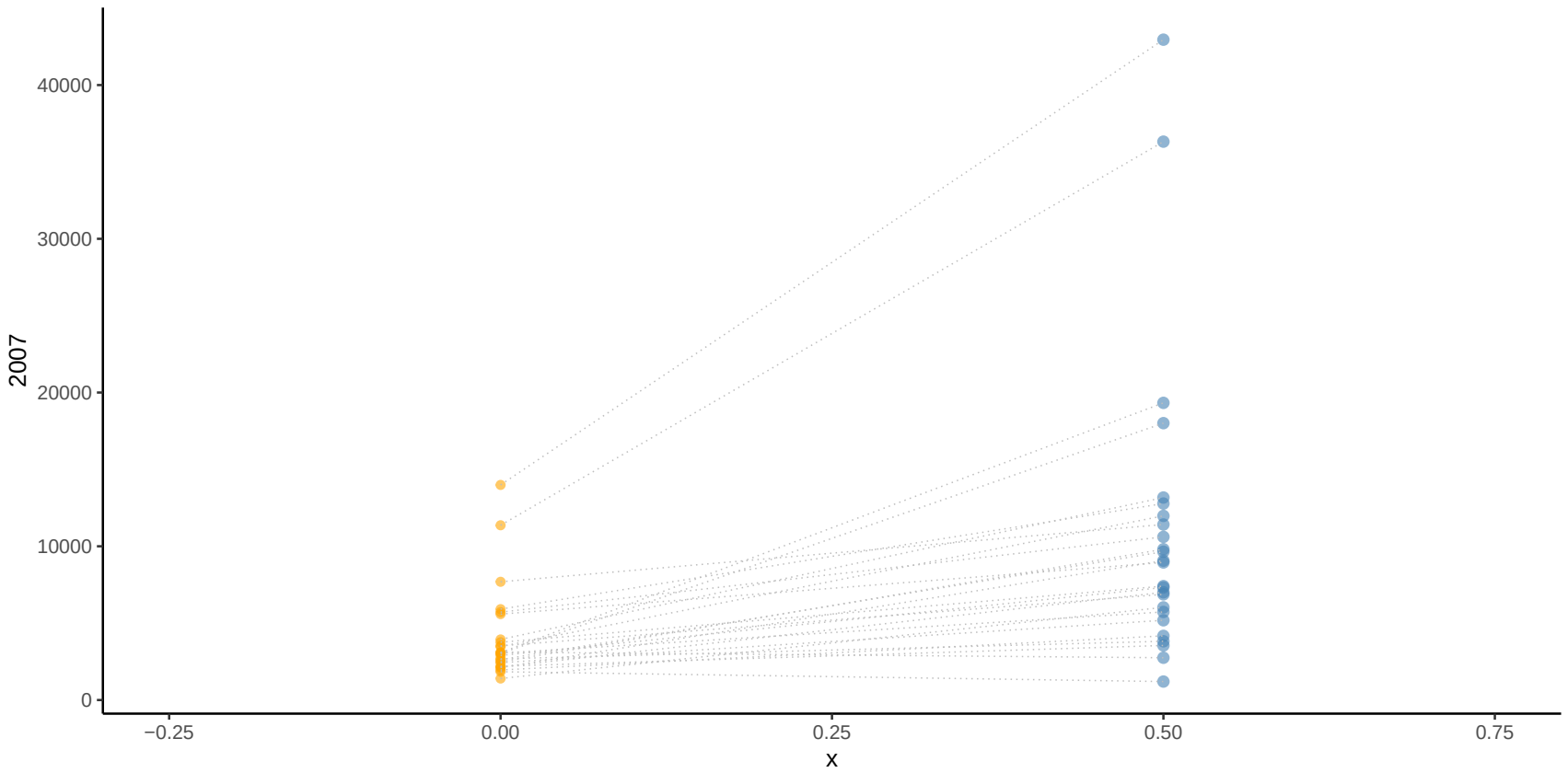


Line graph


gdpPercap
90000
60000
30000
0
1960
1980
2000
year
Norway
Ireland
Switzerland
Netherlands
Austria
Iceland
Bulgaria
Serbia
Montenegro
Albania
Turkey
Bosnia and Herzegovina

Dot plot

dotplot สามารถใช้แสดงพัฒนาการ
หรือการเปลี่ยนแปลงของตัวแปรระหว่างจุดเวลาสองจุดได้



Slope plot