KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 12:
Uji Kualitas Data Melalui Grafik

Outline

- Pembuatan plot yang sesuai dengan tipe data
- Penulisan label pada plot
- Penginterpretasian yang sesuai dengan bentuk dan jenis plot

Plot yang Sesuai untuk Tipe Data

- Pada Materi 4 sudah belajar tentang visualisasi dasar sesuai tipe data masing masing
- Grafik / Plot sangat penting dalam mengenali suatu dataset
- Plot yang tidak sesuai dengan tipe data *doesn't mean anything*
- Statistika deskriptif seringkali tidak cukup untuk mengenal karakteristik data

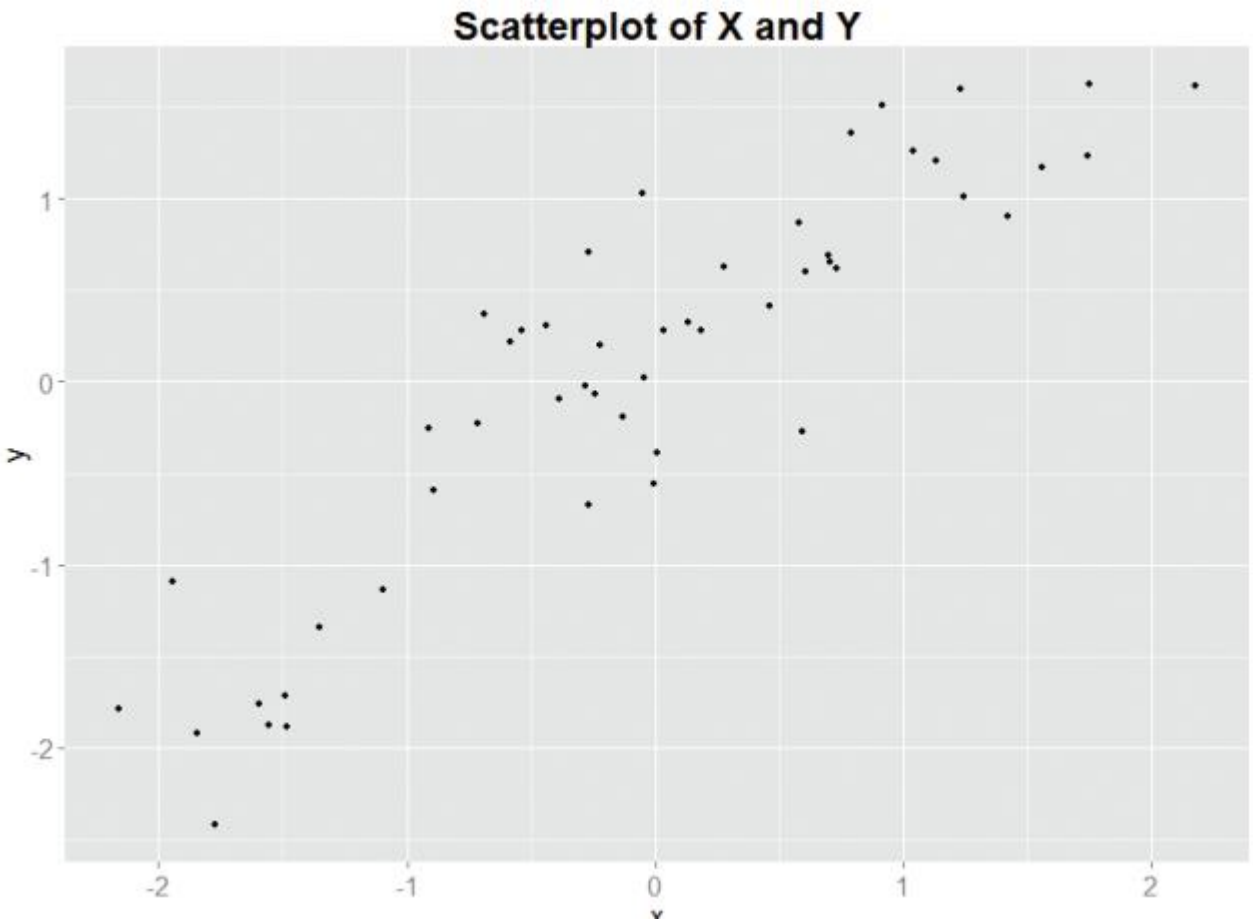
Contoh 1

```
summary(data)
       x                    y           
 Min.   :-1.90483   Min.   :-2.16545  
 1st Qu.:-0.66321   1st Qu.:-0.71451  
 Median : 0.09367   Median :-0.03797  
 Mean   : 0.02522   Mean   :-0.02153  
 3rd Qu.: 0.65414   3rd Qu.: 0.55738  
 Max.   : 2.18471   Max.   : 1.70199  
```

- Melalui statistika deskriptif seperti di samping, cukupkah informasi tersebut?
- Adakah hubungan antara atribut X dan Y?
- Pentingkah variabel X / Y dalam dataset? Seberapa penting?

X and Y, Scattered and Plotted

- Melalui Scatterplot, secara visual dapat melihat hubungan antara X dan Y
- X dan Y yang bertipe *continuous* / data non-diskrit/non-kategorik cocok dengan scatterplot
- Apa hubungan antara X dan Y secara visual? (belum masuk statistika inferensial)



Contoh 2: Anscombe's Quartet

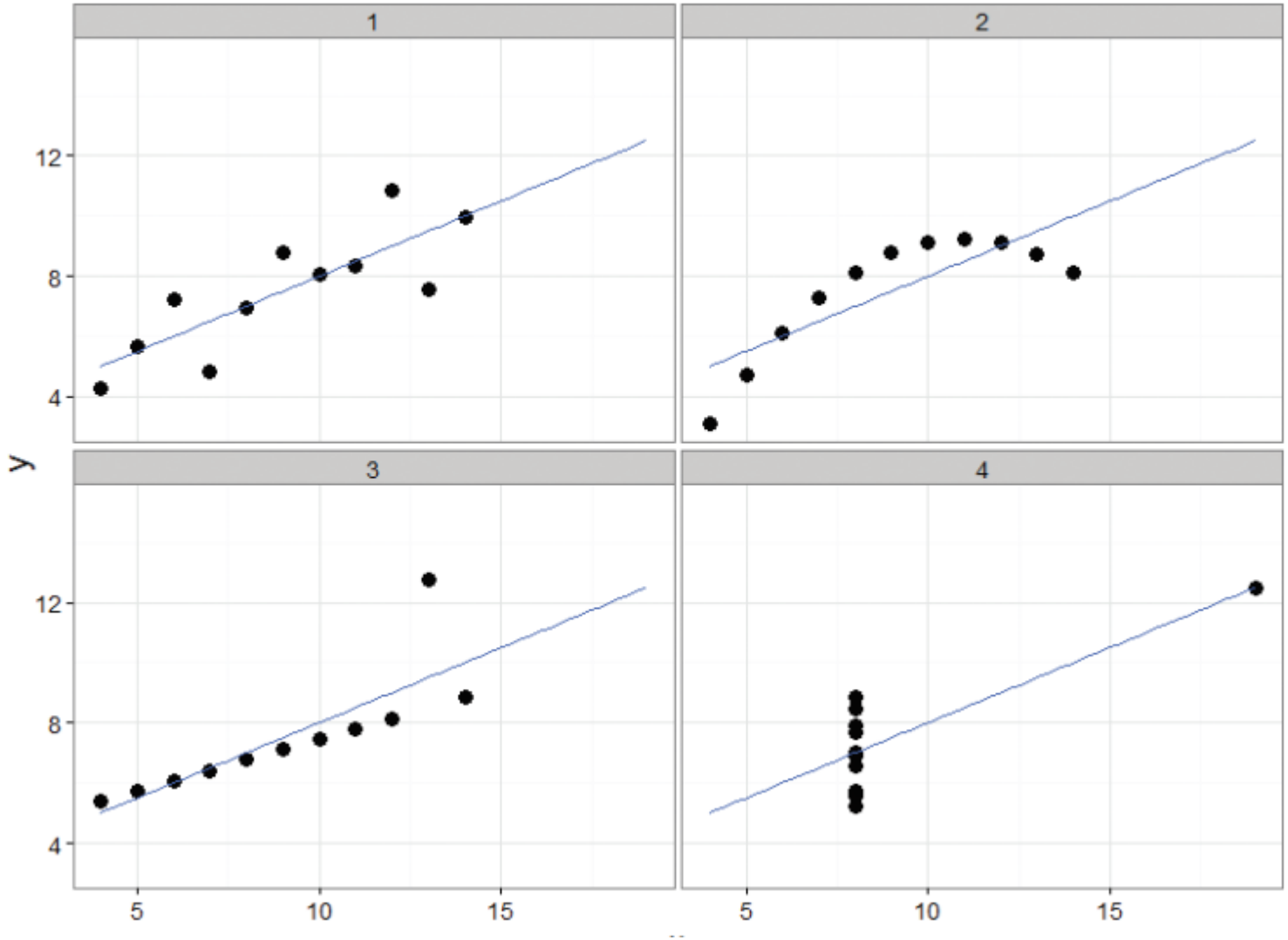
Data oleh statistikawan Perancis, Francis Anscombe (1973)

#1		#2		#3		#4	
x	y	x	y	x	y	x	y
4	4.26	4	3.10	4	5.39	8	5.25
5	5.68	5	4.74	5	5.73	8	5.56
6	7.24	6	6.13	6	6.08	8	5.76
7	4.82	7	7.26	7	6.42	8	6.58
8	6.95	8	8.14	8	6.77	8	6.89
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	7.71
11	8.33	11	9.26	11	7.81	8	7.91
12	10.84	12	9.13	12	8.15	8	8.47
13	7.58	13	8.74	13	12.74	8	8.84
14	9.96	14	8.10	14	8.84	19	12.50

Statistical Property	Value
Mean of *x*	9
Variance of *y*	11
Mean of *y*	7.50 (to 2 decimal points)

Variance of *y*	4.12 or 4.13 (to 2 decimal points)
Correlations between *x* and *y*	0.816
Linear regression line	$y = 3.00 + 0.50x$ (to 2 decimal points)

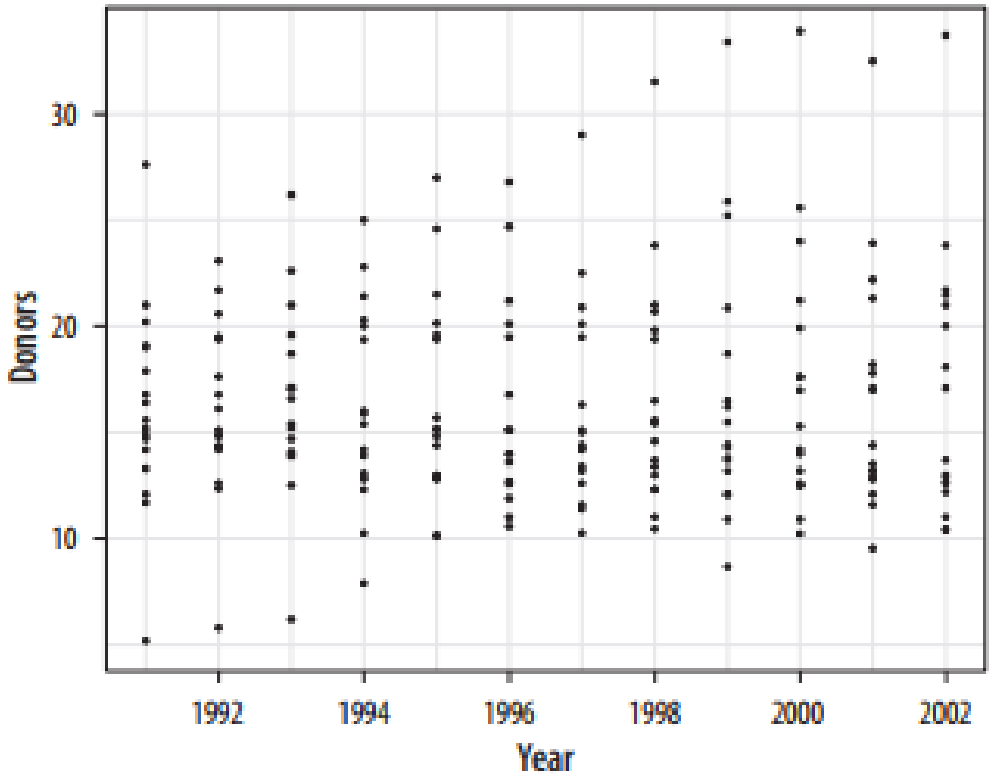
However



Sesuaikah Visualisasi dengan Data?

```
organdata %>% select(1:6) %>% sample_n(size = 10)
```

```
## # A tibble: 10 x 6
##    country        year       donors   pop pop_dens   gdp
##    <chr>          <date>      <dbl> <int>    <dbl> <int>
##  1 Switzerland    NA             NA    NA       NA    NA
##  2 Switzerland    1997-01-01   14.3  7089     17.2  27675
##  3 United Kingdom 1997-01-01   13.4 58283     24.0  22442
##  4 Sweden         NA             NA  8559      1.90 18660
##  5 Ireland        2002-01-01   21.0  3932      5.60 32571
##  6 Germany        1998-01-01   13.4 82047     23.0  23283
##  7 Italy          NA             NA 56719     18.8  17430
##  8 Italy          2001-01-01   17.1 57894     19.2  25359
##  9 France         1998-01-01   16.5 58398     10.6  24044
## 10 Spain          1995-01-01   27.0 39223      7.75 15720
```



How about this plot?

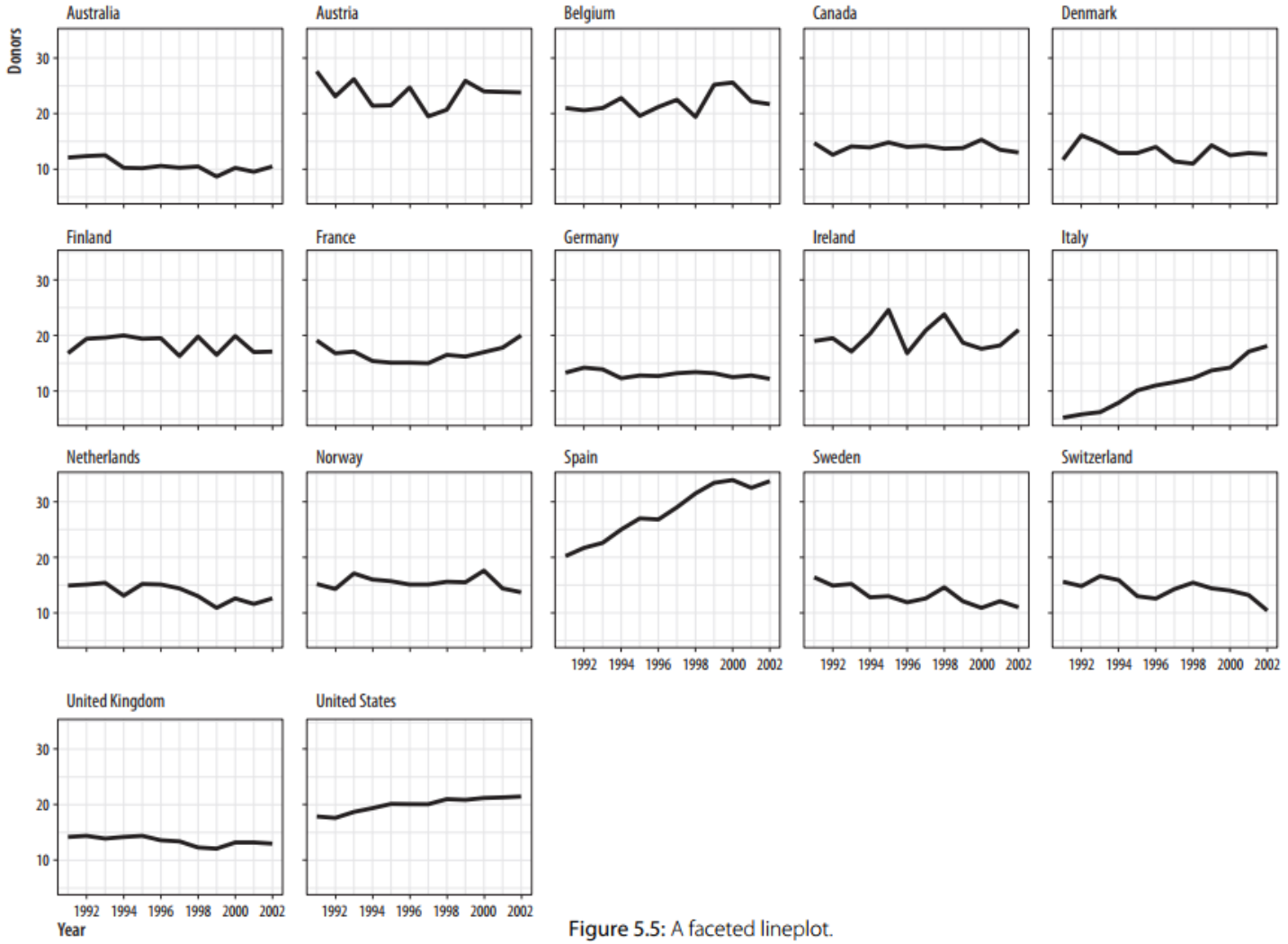
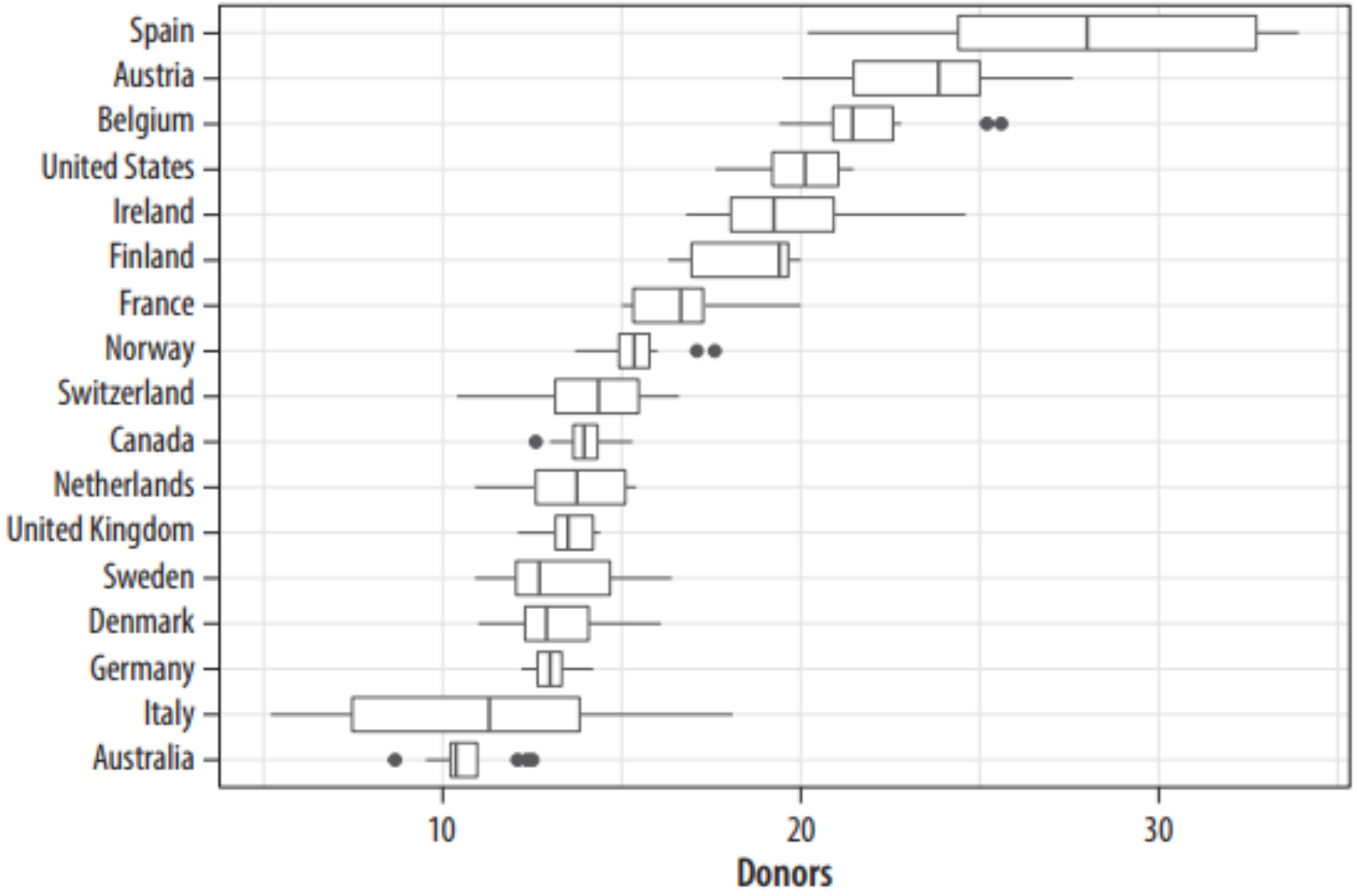


Figure 5.5: A faceted lineplot.

Bagaimana Jika

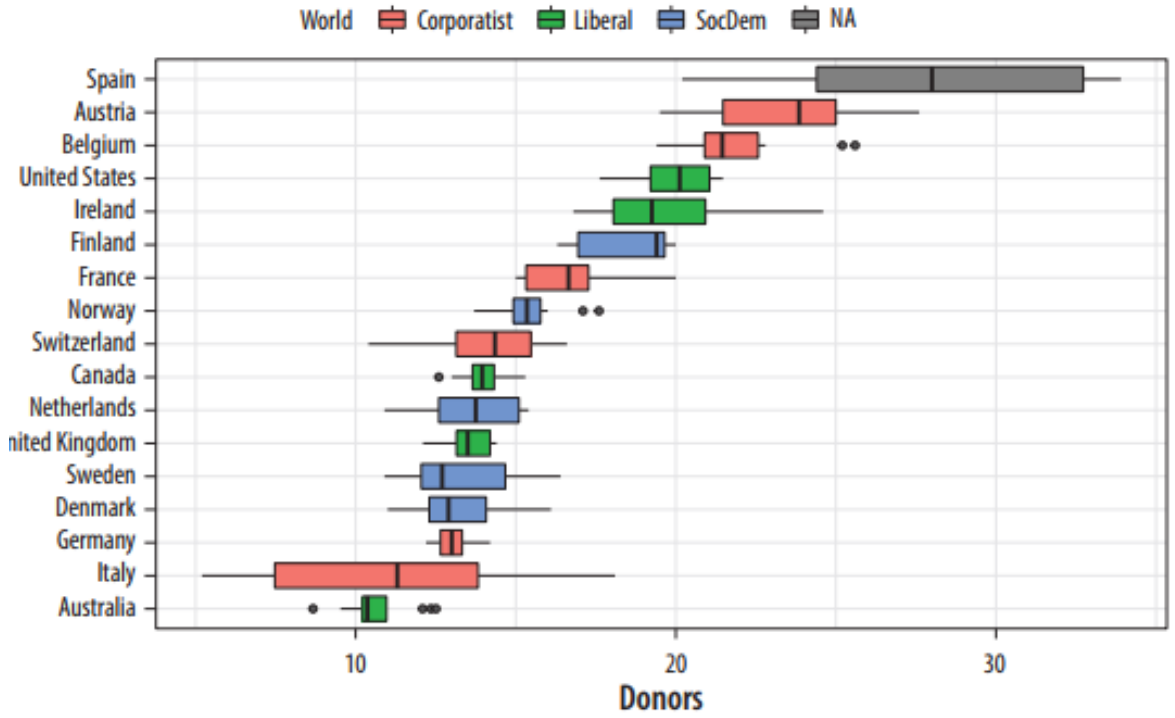
- Stakeholder ingin mengetahui negara mana yang mempunyai donor terbanyak dibanding yang lainnya?
- Negara mana yang mempunyai kecenderungan donor lebih tinggi dari negara lain?
- Plot apa yang digunakan?

Boxplot, Ordered by Mean



Penulisan Label pada Plot

- Seringkali sebuah plot hanya mencakup 2 Atribut (2 Dimensi)
- Penambahan warna pada plot dapat menambah atribut informasi pada plot

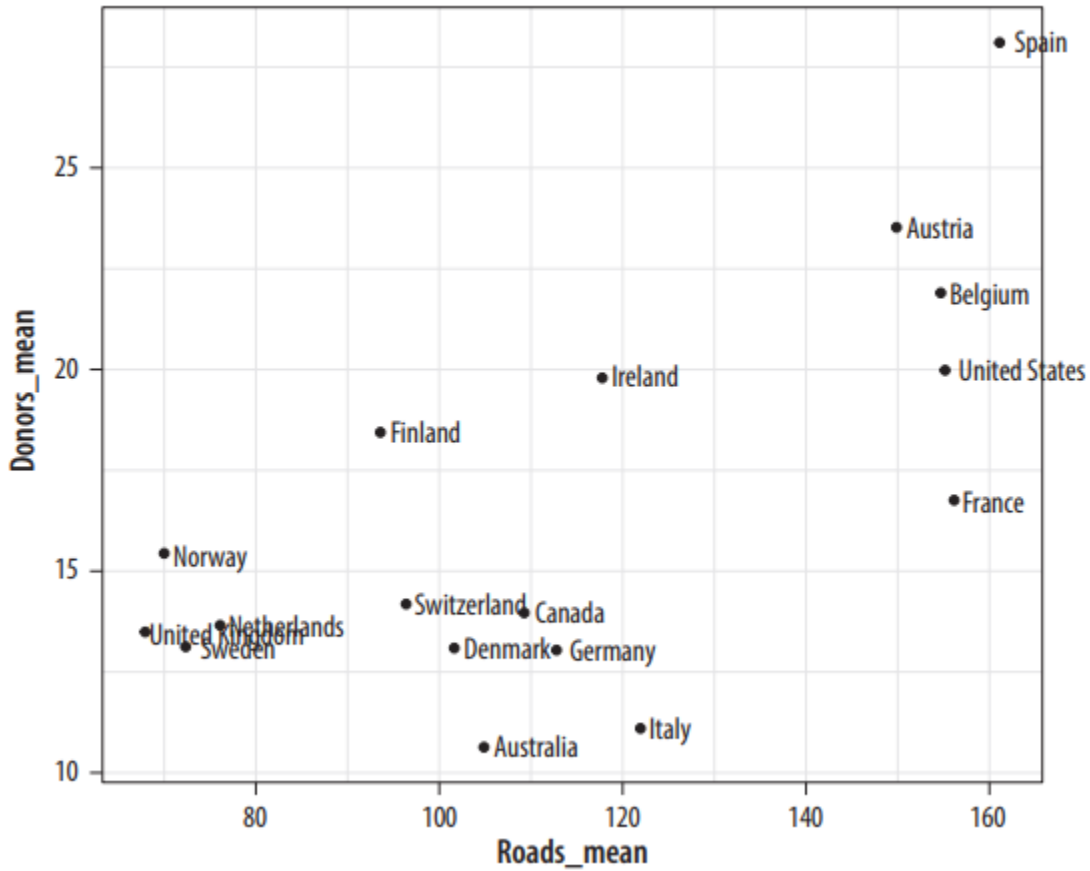


Plot dengan direct text

- Dengan menambahkan geom_text setelah geom_point sebagai plot baru, dan atribut hjust untuk penambahan jarak text dengan point

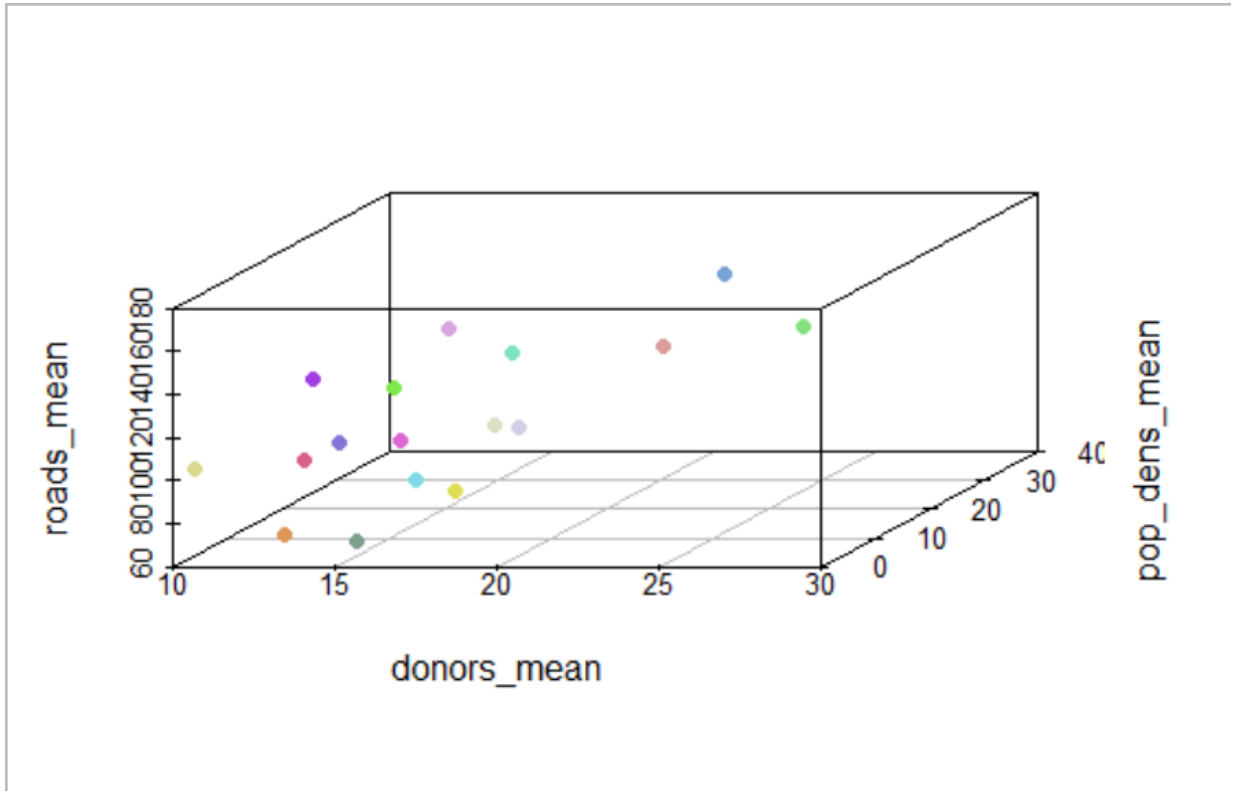
```
by_country ← organdata %>% group_by(consent_law, country) %>%
    summarize_if(is.numeric, funs(mean, sd), na.rm = TRUE) %>%
    ungroup()
```

```
p ← ggplot(data = by_country,
           mapping = aes(x = roads_mean, y = donors_mean))

p + geom_point() + geom_text(mapping = aes(label = country), hjust = 0)
```



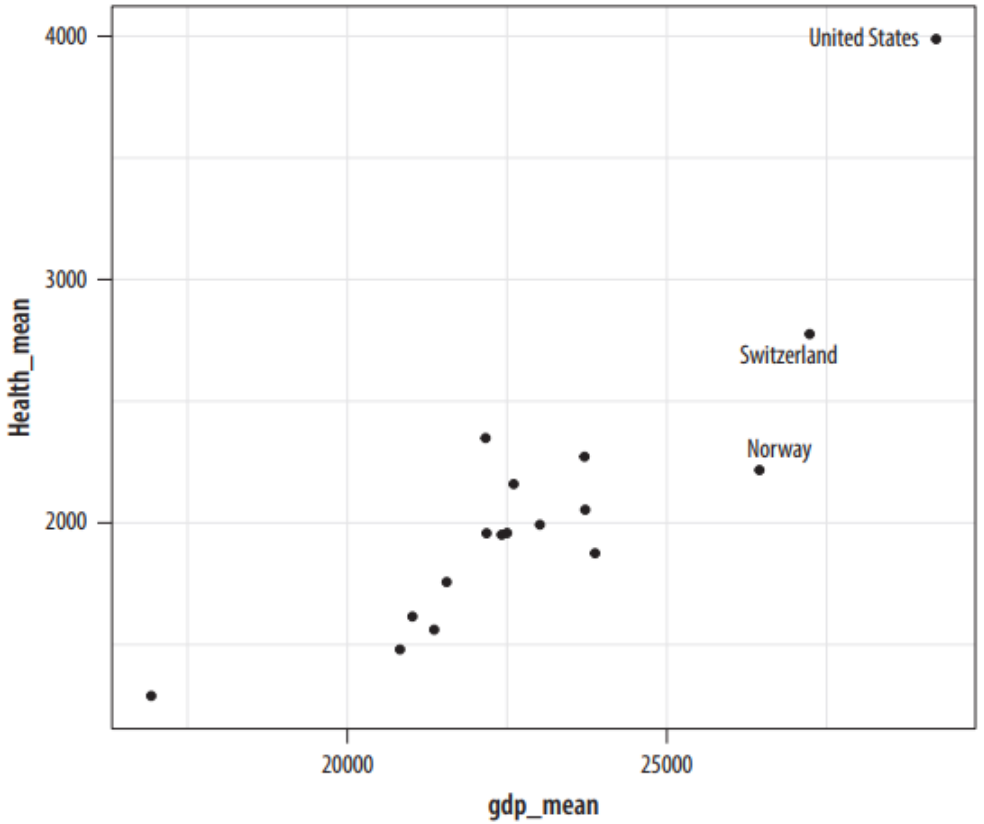
3D plot?

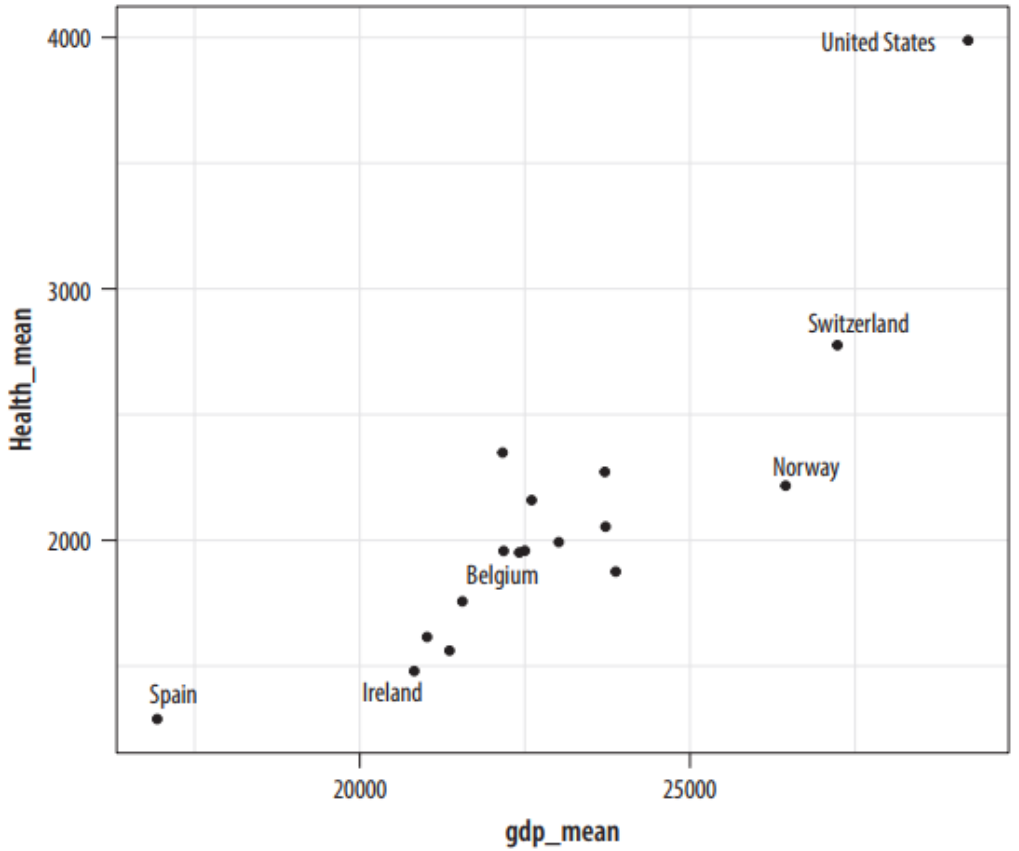
- Sebuah package scatterplot3d dapat menunjukkan scatterplot dengan atribut / axis lebih dari 2 (X,Y,Z) atau bisa disebut 3 dimensi



Labelling point of interest

- Seringkali, terlalu banyak label mengakibatkan plot sulit dibaca
- Ada kalanya label ditulis untuk *highlighting insight*



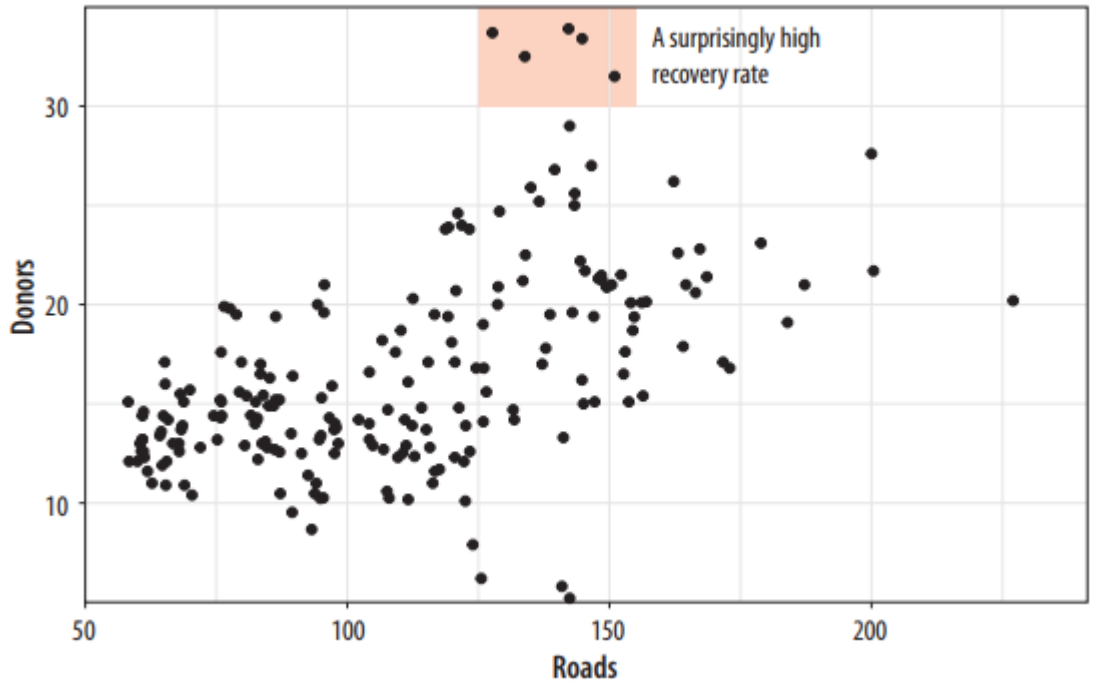


```
p ← ggplot(data = by_country,
           mapping = aes(x = gdp_mean, y = health_mean))

p + geom_point() +
    geom_text_repel(data = subset(by_country, gdp_mean > 25000),
                    mapping = aes(label = country))

p ← ggplot(data = by_country,
           mapping = aes(x = gdp_mean, y = health_mean))

p + geom_point() +
    geom_text_repel(data = subset(by_country,
                                  gdp_mean > 25000 | health_mean < 1500 |
                                  country %in% "Belgium"),
                    mapping = aes(label = country))
```

Labelling point of interest (continued)

- Bisa juga menggunakan fungsi annotate untuk memberi teks tambahan pada point of interest / highlight
- Dan memberikan warna khusus untuk pembeda

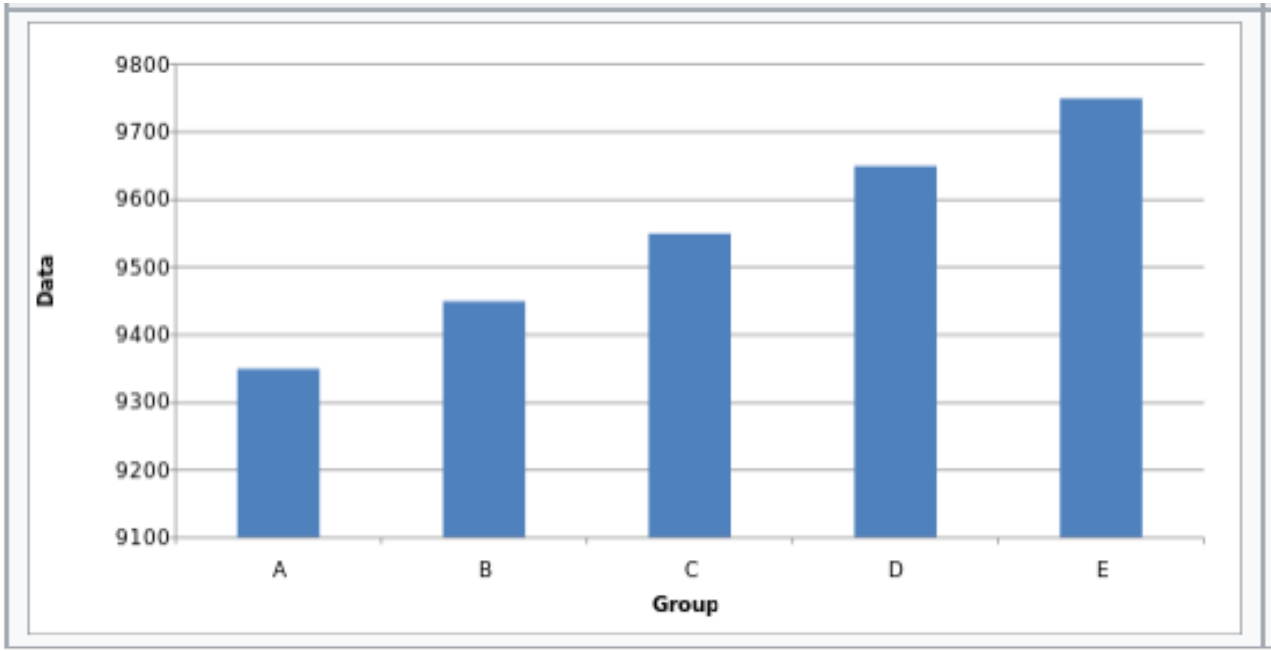
```
p ← ggplot(data = organdata,
           mapping = aes(x = roads, y = donors))
p + geom_point() +
    annotate(geom = "rect", xmin = 125, xmax = 155,
             ymin = 30, ymax = 35, fill = "red", alpha = 0.2) +
    annotate(geom = "text", x = 157, y = 33,
             label = "A surprisingly high \n recovery rate.", hjust = 0)
```

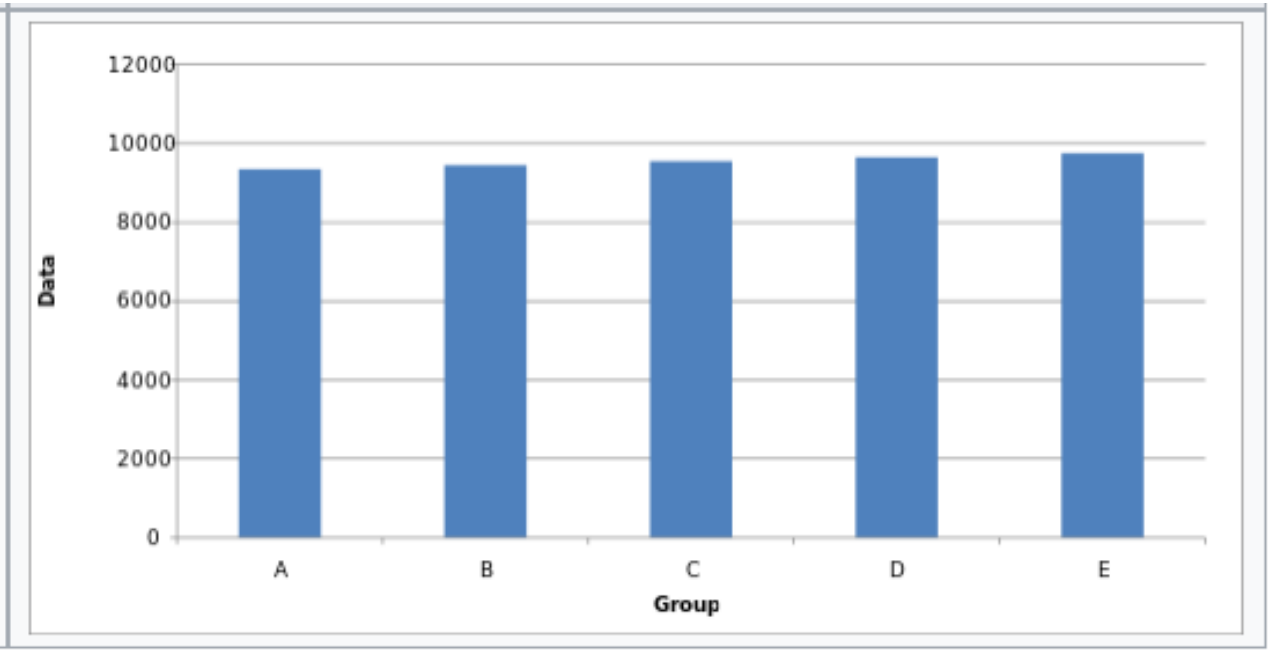


Labelling and Using Appropriate Plots

- Dapat mengenal karakteristik data dengan lebih dalam
- Dapat mengkomunikasikan karakteristik data dengan lebih mudah
- Sebagai langkah *Exploratory* sebelum melanjutkan ke analisis inferensial dan/atau *Machine Learning*
- Mengetahui anomali data meskipun secara deskriptif terlihat normal (Anscombe's Quartet)

Plots can Lie

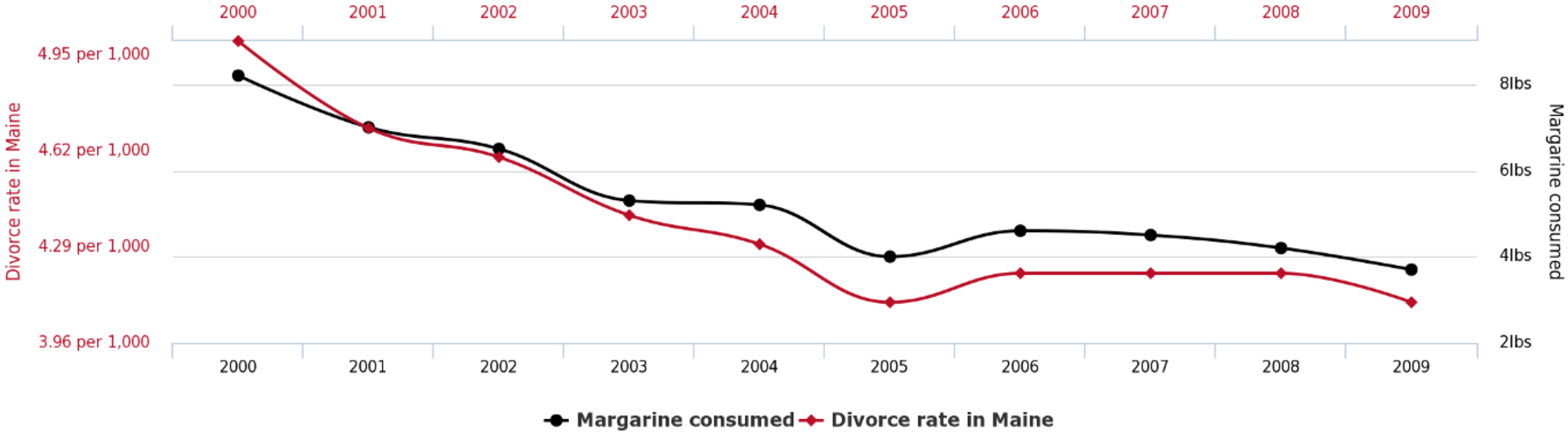




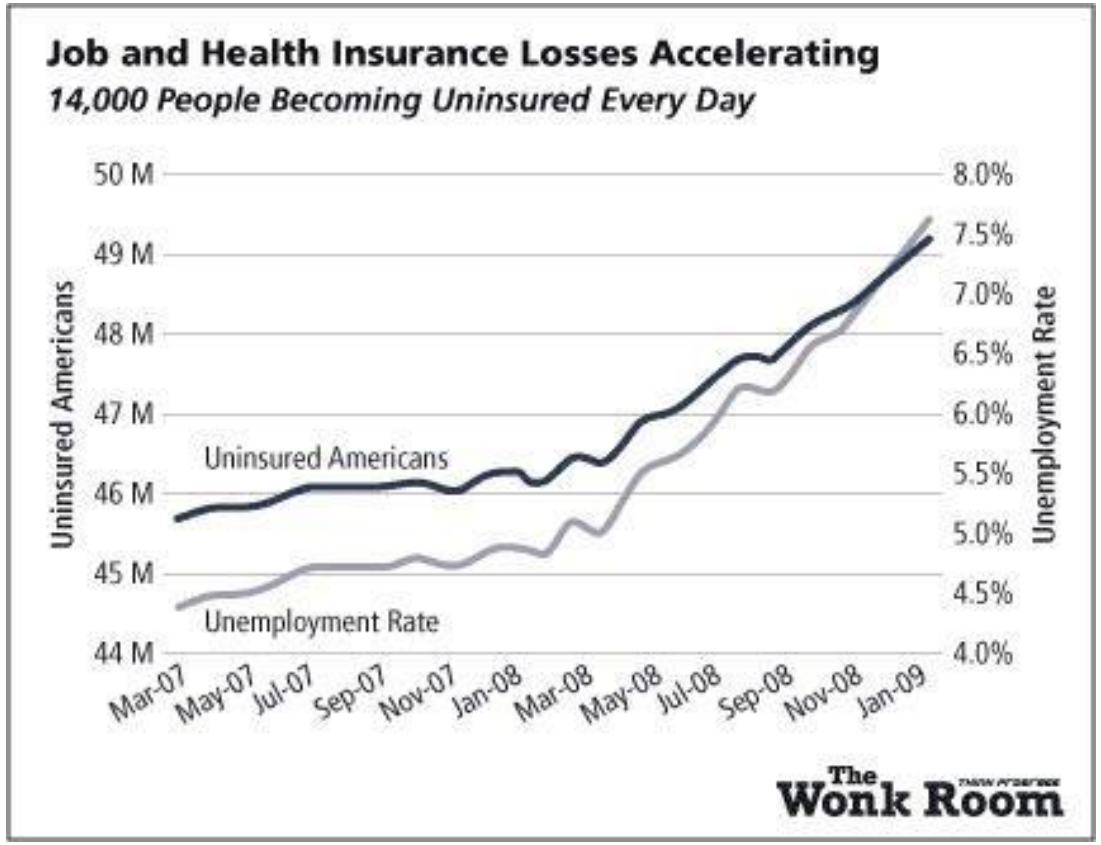
Divorce rate in Maine

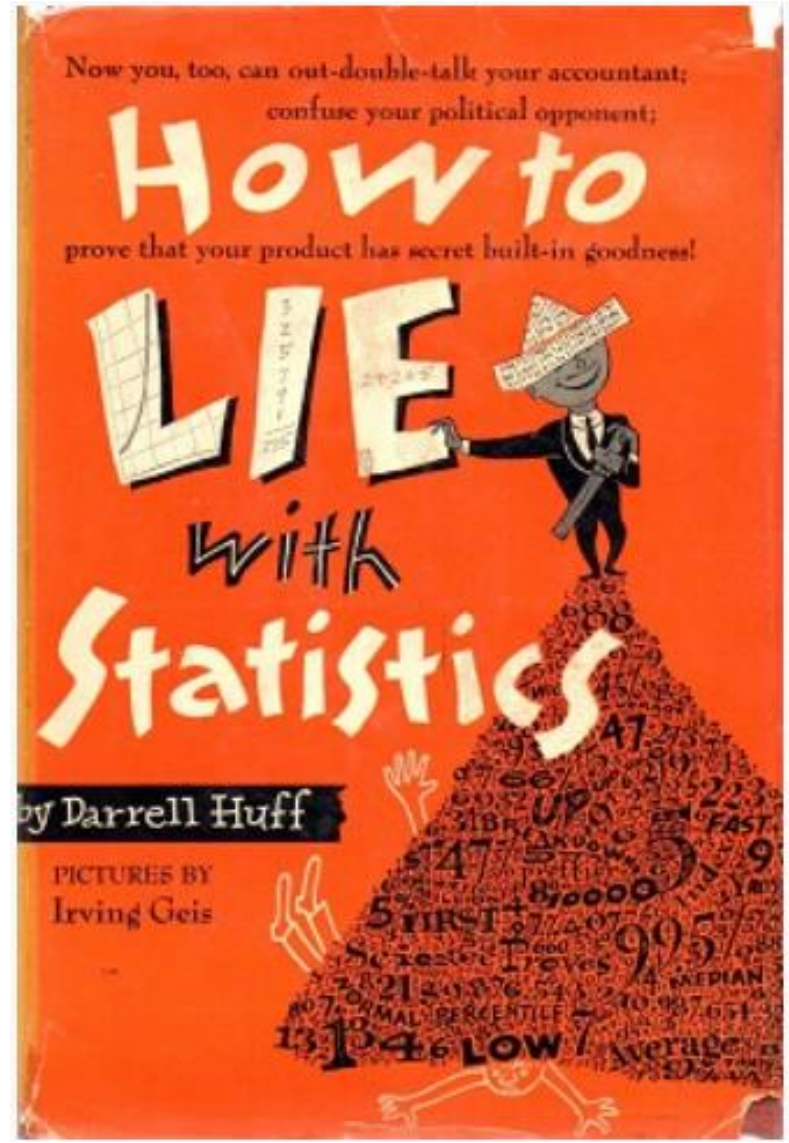
correlates with

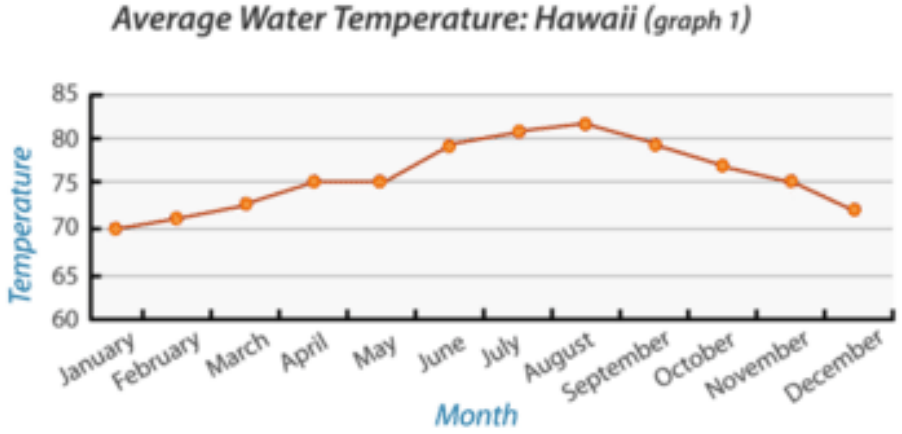
Per capita consumption of margarine

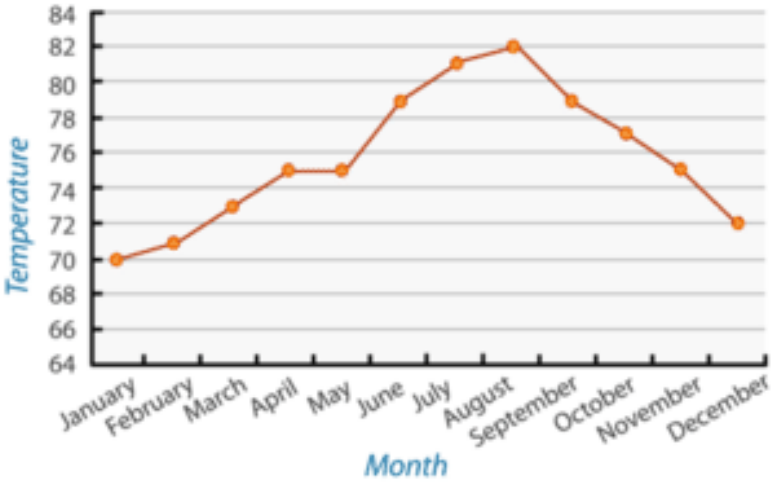


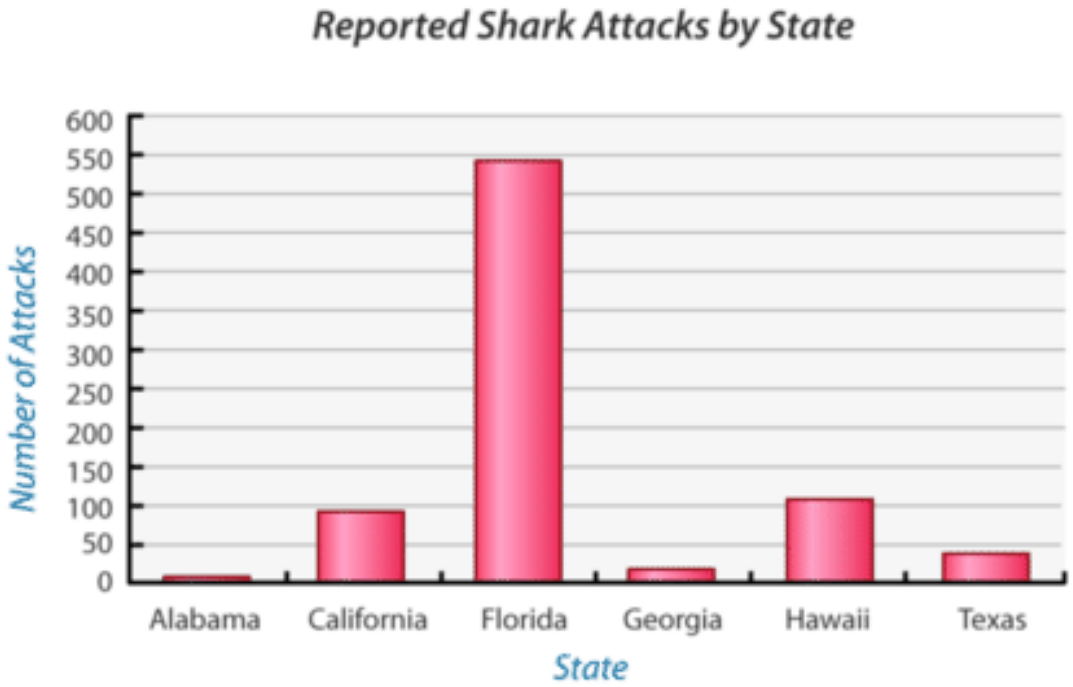
tylervigen.com


Job and Health Insurance Losses Accelerating
14,000 People Becoming Uninsured Every Day
50 M
49 M
48 M
47 M
46 M
45 M
44 M
Uninsured Americans
8.0%
7.5%
7.0%
6.5%
6.0%
5.5%
5.0%
4.5%
4.0%
Unemployment Rate
Uninsured Americans
Unemployment Rate
Mar-07
May-07
Jul-07
Sep-07
Nov-07
Jan-08
Mar-08
May-08
Jul-08
Sep-08
Nov-08
Jan-09
The Wonk Room


Now you, too, can out-double-talk your accountant;
confuse your political opponent;
How to
prove that your product has secret built-in goodness!
LIE
with
Statistics
by Darrell Huff
PICTURES BY
Irving Geis


Temperature
85
80
75
70
65
60
January
February
March
April
May
June
July
August
September
October
November
December
Month


Temperature
84
82
80
78
76
74
72
70
68
66
64
January
February
March
April
May
June
July
August
September
October
November
December
Month


Reported Shark Attacks by State
Number of Attacks
600
550
500
450
400
350
300
250
200
150
100
50
0
Alabama
California
Florida
Georgia
Hawaii
Texas
State


Have You Ate Pizza This Week?
120
110
100
90
80
70
60
50
40
30
20
10
0
Yes 84%
No 16%

Conclusion

- Interpretasi dapat terpengaruh dari plot yang digunakan
- Plotting yang tidak sesuai dapat menyembunyikan informasi dan berakibat buruk pada pengenalan karakteristik (ada karakteristik yg tersembunyi)
- Plotting yang disengaja dapat digunakan untuk berbohong / exaggeration
- Visualisasi untuk presentasi / storytelling dapat dimanipulasi
- Untuk EDA, visualisasi yang dimanipulasi akan merugikan diri sendiri.
- Kasus lie with statistics sering ditemui di sekitar kita.

Tugas

- ✓ Praktikum: Membuat plot yang sesuai tipe data dari dataset organdata (package socviz):
  a) Melihat tren donor dari 3 negara berbeda
  b) Melihat ranking negara dengan jumlah pendonor terbanyak dari tahun 1991 sampai 2002
  c) Melihat interaksi variabel populasi dengan jumlah pendonor di 3 negara berbeda
- ✓ Menyimpulkan hasil plot (mendapatkan insight dari plot) dari 3 tugas di atas.
- ✓ Laporkan dalam bentuk laporan praktikum dengan menyertakan langkah langkah pengerjaan berupa narasi dan screenshot R serta hasil analisis dari setiap langkah.

- ○ Tugas dikerjakan berkelompok.
- ○ Tugas dikumpulkan paling lambat pukul 23.59 WIB di LMS.
- ○ Beri nama file tugas: Tugas 12_Kelompok XX. (Contoh: Tugas 12_Kelompok 01)

FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN

PERTUKARAN MAHASISWA merdeka

Kampus Merdeka INDONESIA JAYA

Terima Kasih

SMART UNIVERSITY

TOP #465 WORLD UNIVERSITY RANKINGS UNAIR UNGGUL

ZONA INTEGRITAS menuju wilayah bebas korupsi