Visualisasi

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

October 10, 2019

Selayang Pandang

1. Visualisasi Data
2. Tipe-tipe Grafik
3. Visualisasi Efektif

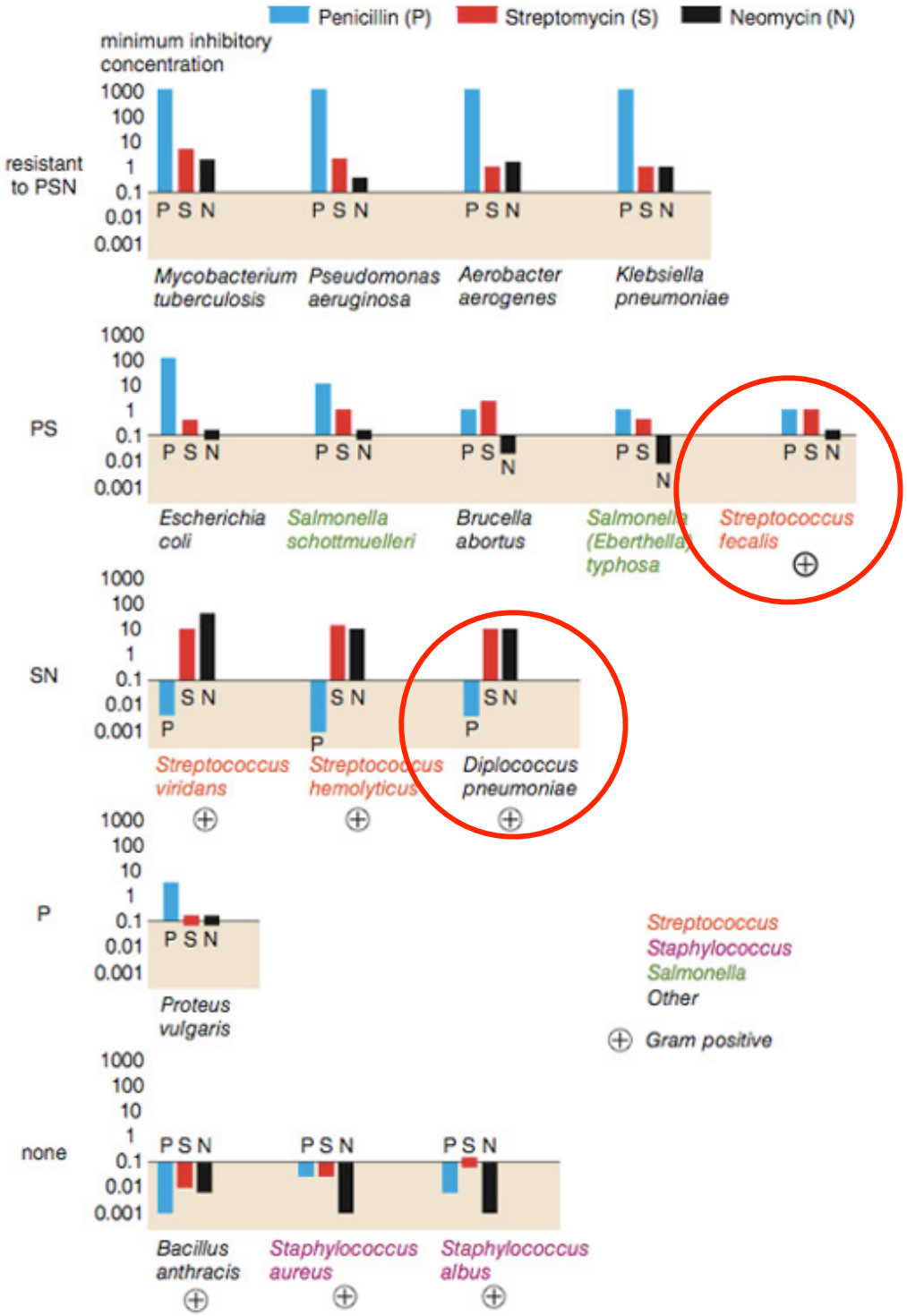
Bahan Bacaan

1. Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Analytics Press.
2. `https://nces.ed.gov/forum/pdf/NCES_table_design.pdf`

"There are three kinds of lies:
lies, damned lies, and statistics."

Visualisasi Data

Sebagai analis data yang baik,
penting sekali untuk memvisualisasikan data.
Mengapa?



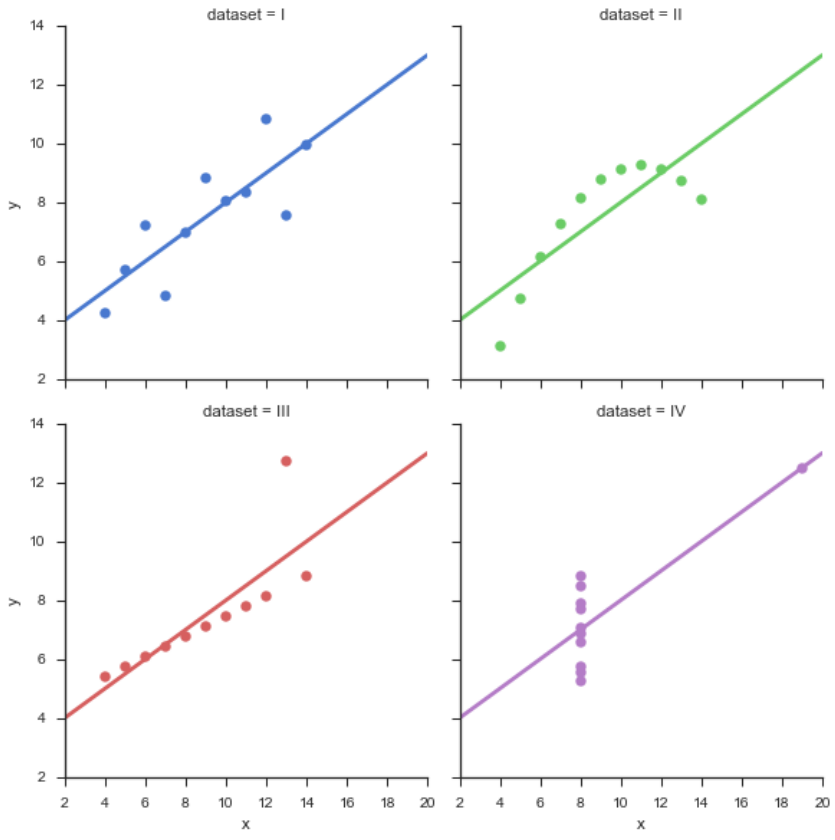
How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

Visualisasi vs Summary Statistics

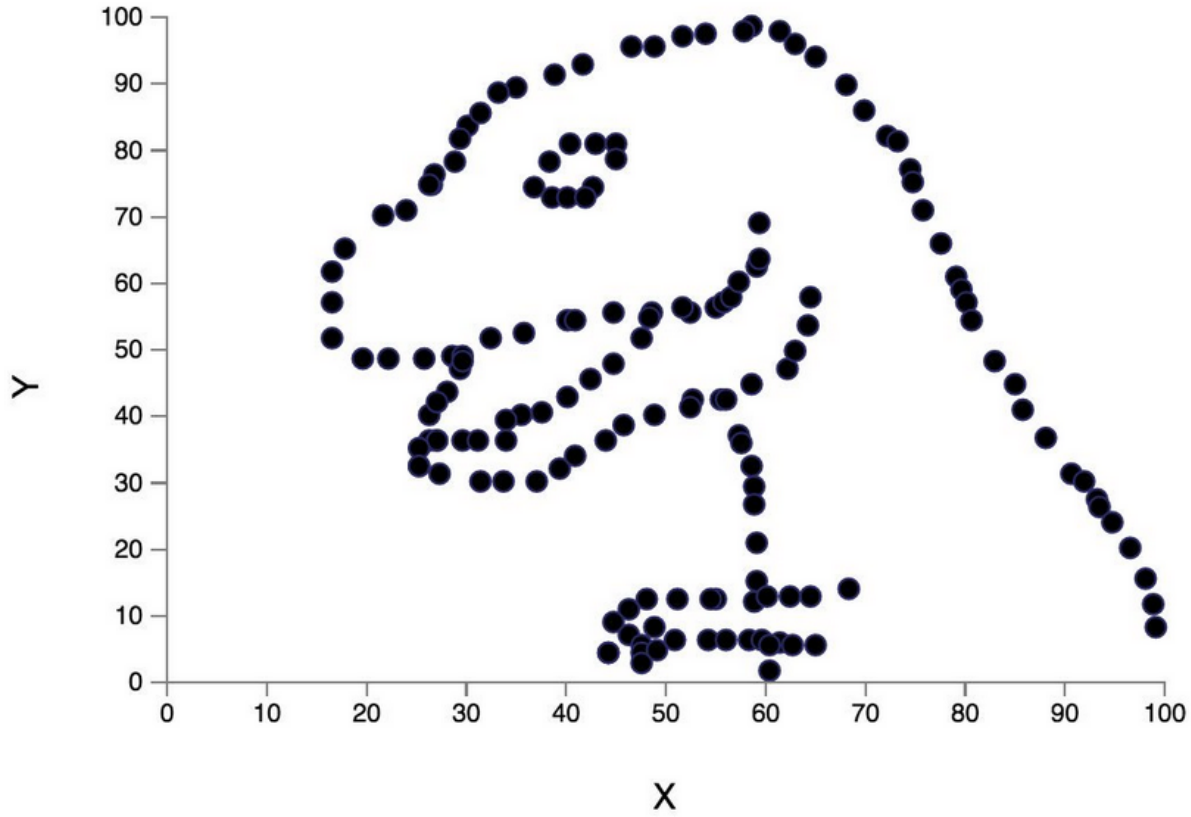


Gambar: *The infamous Anscombe's quartet*

Keempat data tersebut punya
$\mu_x, \mu_y, \sigma_x, \sigma_y, N, \rho$ yang sama!

Don't trust summary statistics.
Always visualize your data!

N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;
Pearson correlation = -0.1772

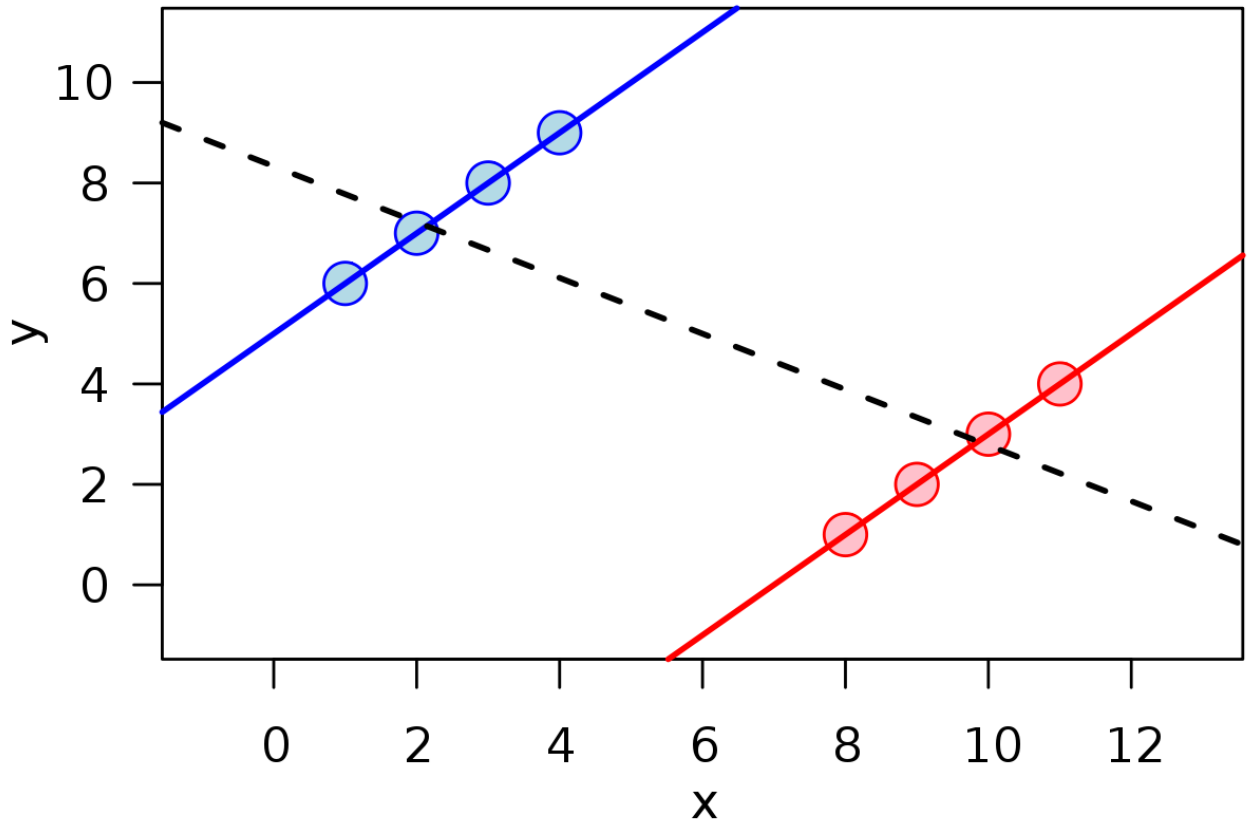


Alberto Cairo @albertocairo · 15 Aug 2016

Don't trust summary statistics. Always visualize your data first robertgrantstats.co.uk/drawmydata.html pic.twitter.com/5j94Dw9UAf

13 895 984

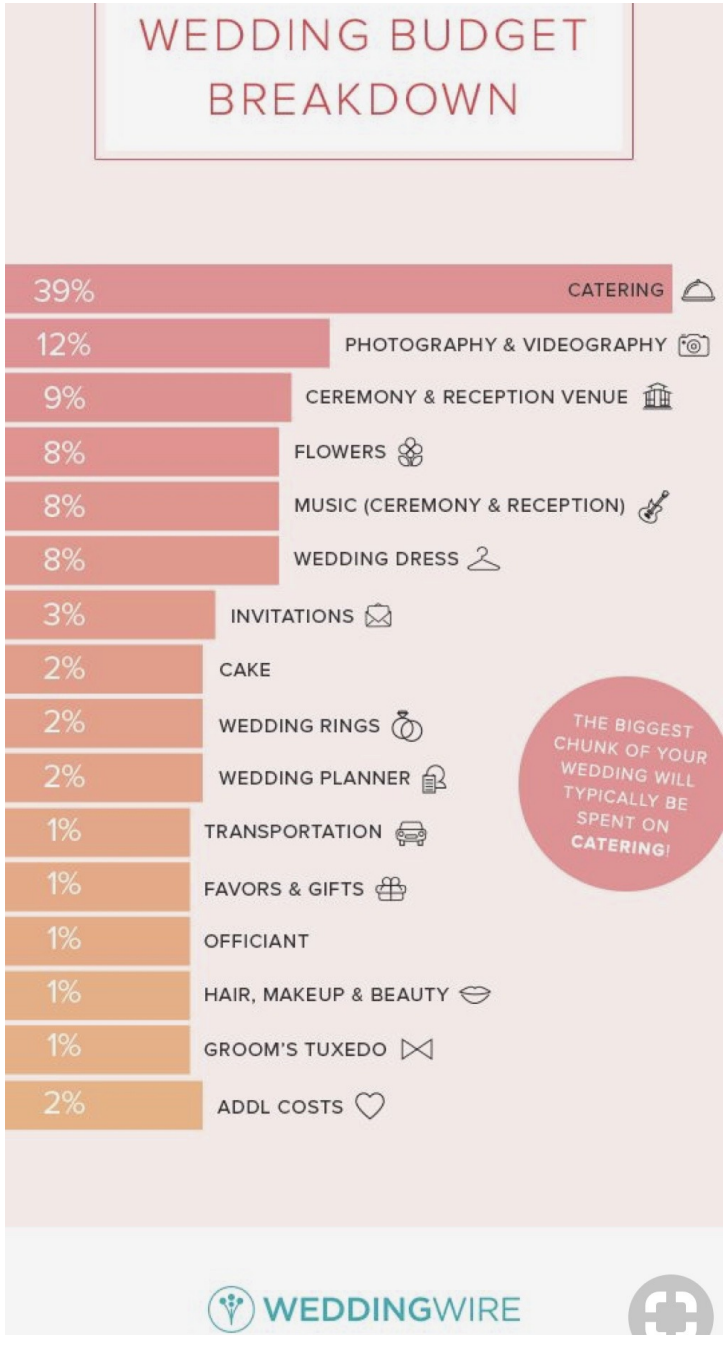
Simpson's Paradox

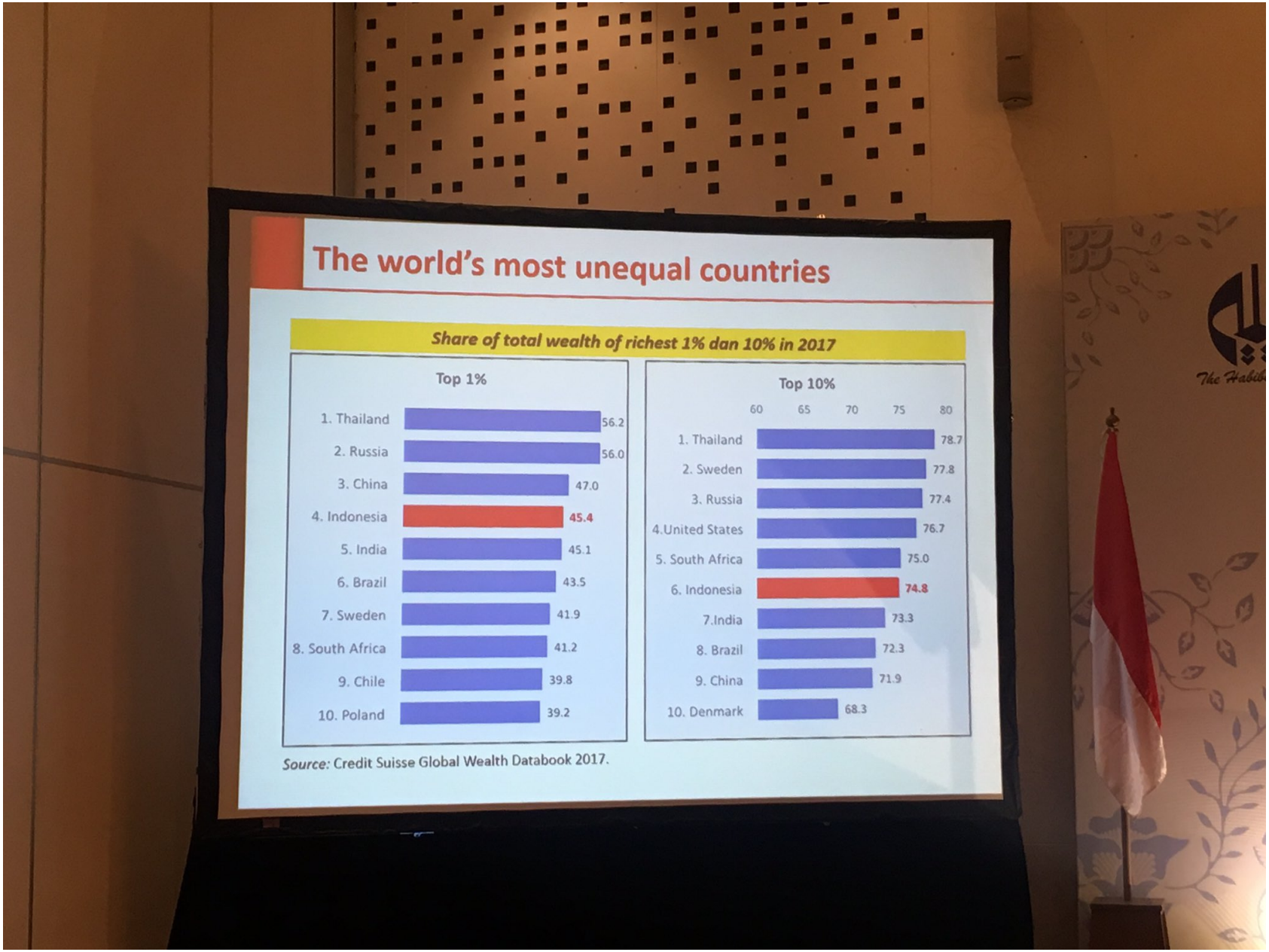


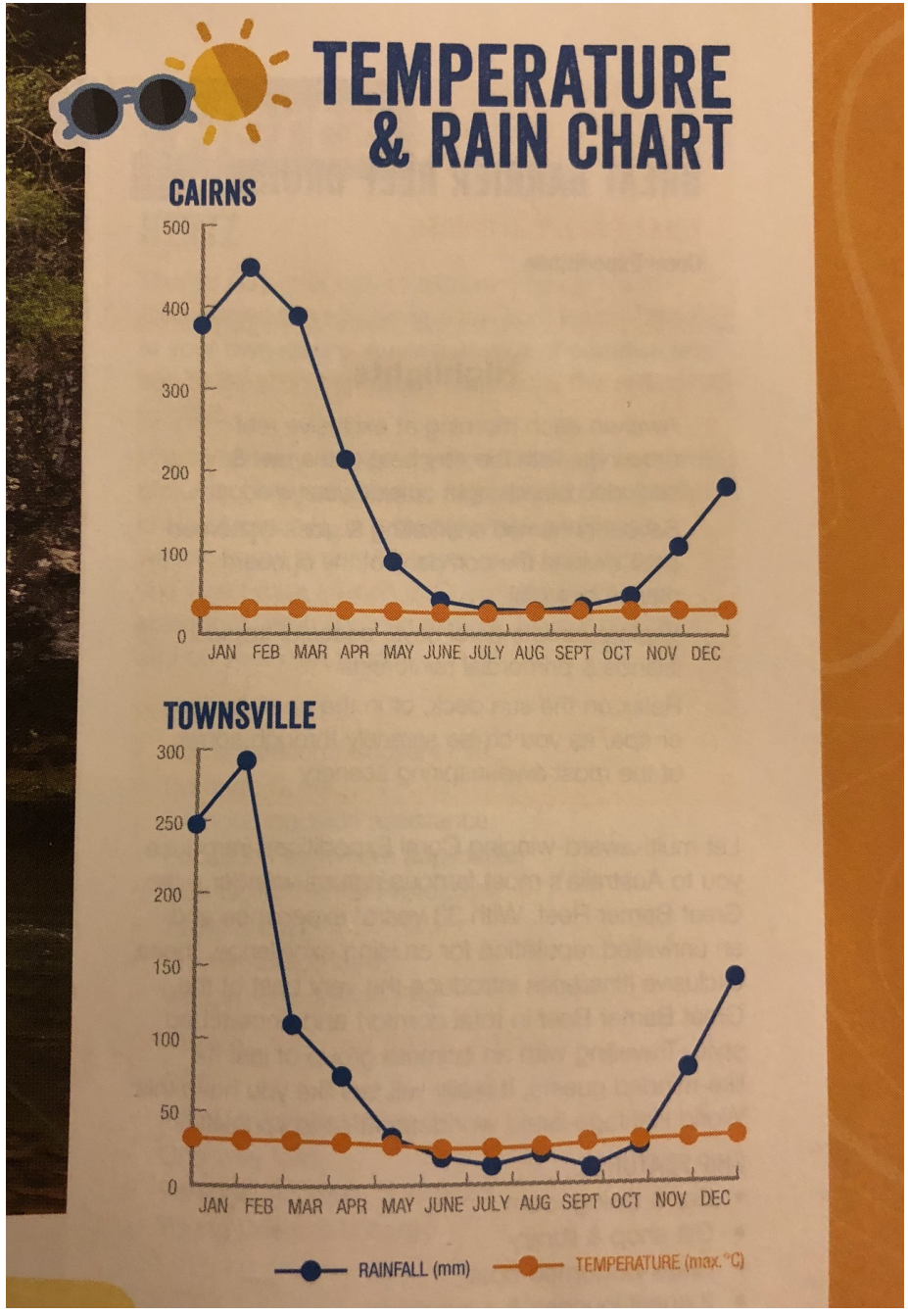
Gambar: Kesalahan dalam interpretasi pola karena tidak dibagi per grup

Dengan visualisasi, bukan berarti bebas masalah.

Kita tetap bisa berbohong!


WEDDING BUDGET BREAKDOWN
39% CATERING
12% PHOTOGRAPHY & VIDEOGRAPHY
9% CEREMONY & RECEPTION VENUE
8% FLOWERS
8% MUSIC (CEREMONY & RECEPTION)
8% WEDDING DRESS
3% INVITATIONS
2% CAKE
2% WEDDING RINGS
2% WEDDING PLANNER
1% TRANSPORTATION
1% FAVORS & GIFTS
1% OFFICIANT
1% HAIR, MAKEUP & BEAUTY
1% GROOM'S TUXEDO
2% ADDL COSTS
THE BIGGEST CHUNK OF YOUR WEDDING WILL TYPICALLY BE SPENT ON CATERING!
WEDDINGWIRE


The world's most unequal countries
Share of total wealth of richest 1% dan 10% in 2017
Top 1%
1. Thailand 56.2
2. Russia 56.0
3. China 47.0
4. Indonesia 45.4
5. India 45.1
6. Brazil 43.5
7. Sweden 41.9
8. South Africa 41.2
9. Chile 39.8
10. Poland 39.2
Top 10%
60 65 70 75 80
1. Thailand 78.7
2. Sweden 77.8
3. Russia 77.4
4.United States 76.7
5. South Africa 75.0
6. Indonesia 74.8
7.India 73.3
8. Brazil 72.3
9. China 71.9
10. Denmark 68.3
Source: Credit Suisse Global Wealth Databook 2017.


TEMPERATURE & RAIN CHART
CAIRNS
500
400
300
200
100
0
JAN FEB MAR APR MAY JUNE JULY AUG SEPT OCT NOV DEC
TOWNSVILLE
300
250
200
150
100
50
0
JAN FEB MAR APR MAY JUNE JULY AUG SEPT OCT NOV DEC
RAINFALL (mm)
TEMPERATURE (max.°C)

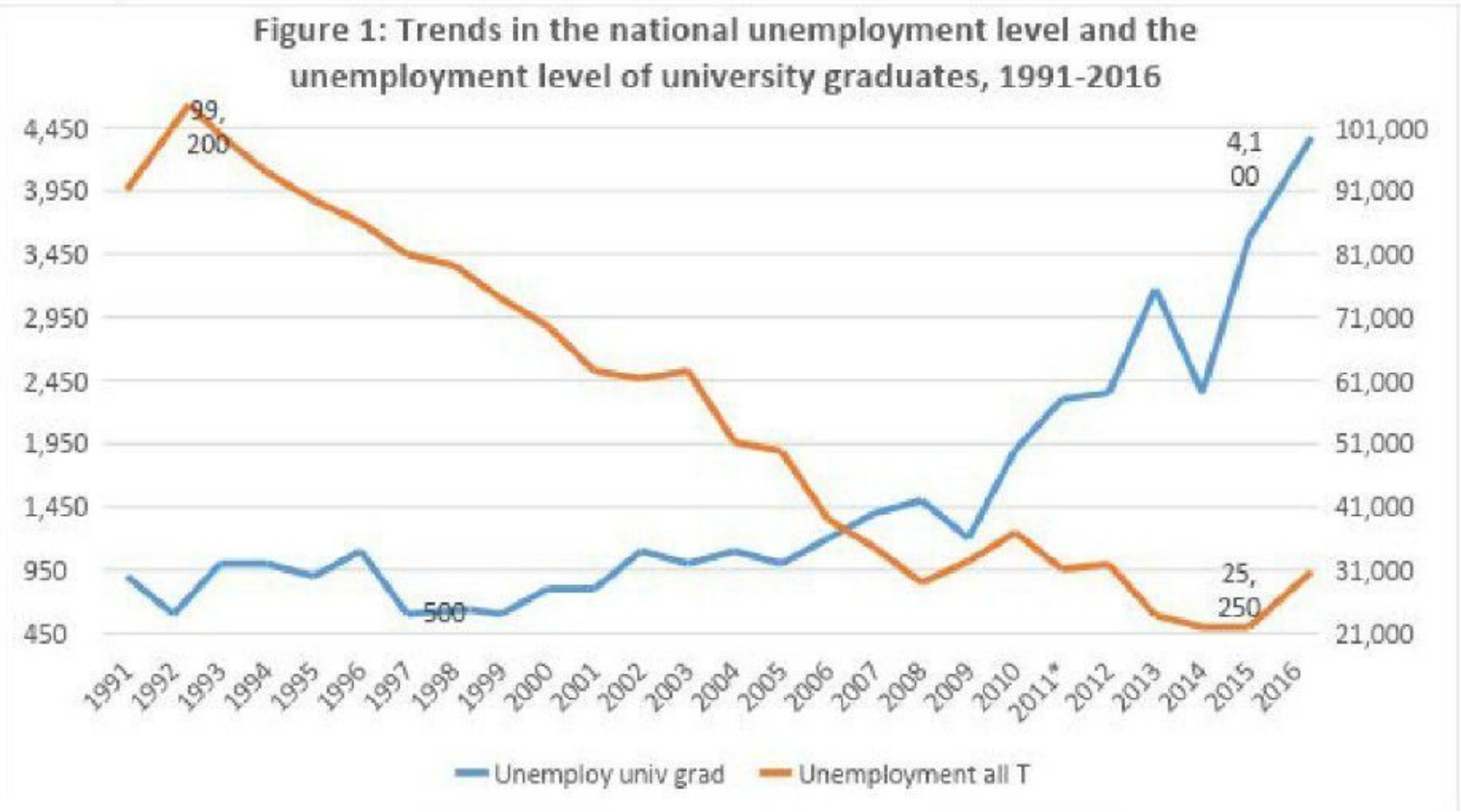
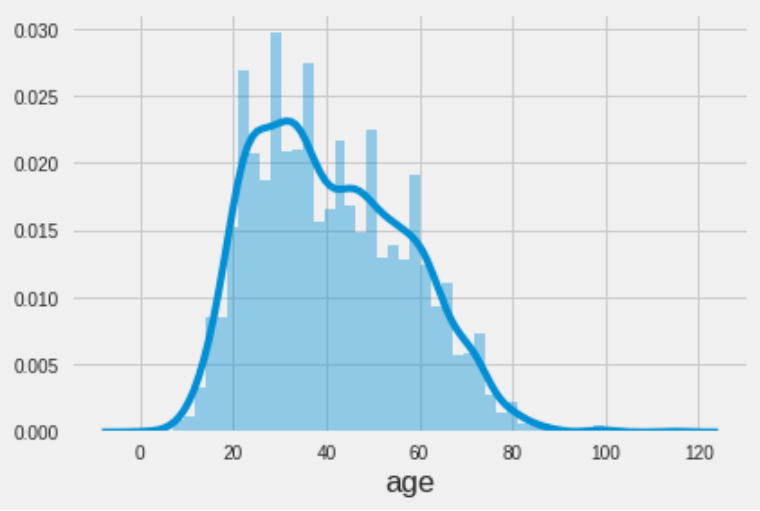


Figure 1: Trends in the national unemployment level and the unemployment level of university graduates, 1991-2016
4,450
3,950
3,450
2,950
2,450
1,950
1,450
950
450
101,000
91,000
81,000
71,000
61,000
51,000
41,000
31,000
21,000
99,200
500
4,100
25,250
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011*
2012
2013
2014
2015
2016
Unemploy univ grad
Unemployment all T

Gambar: Dapat digunakan untuk keuntungan politik. Sumber: @ainunnajib

Tipe-tipe Grafik

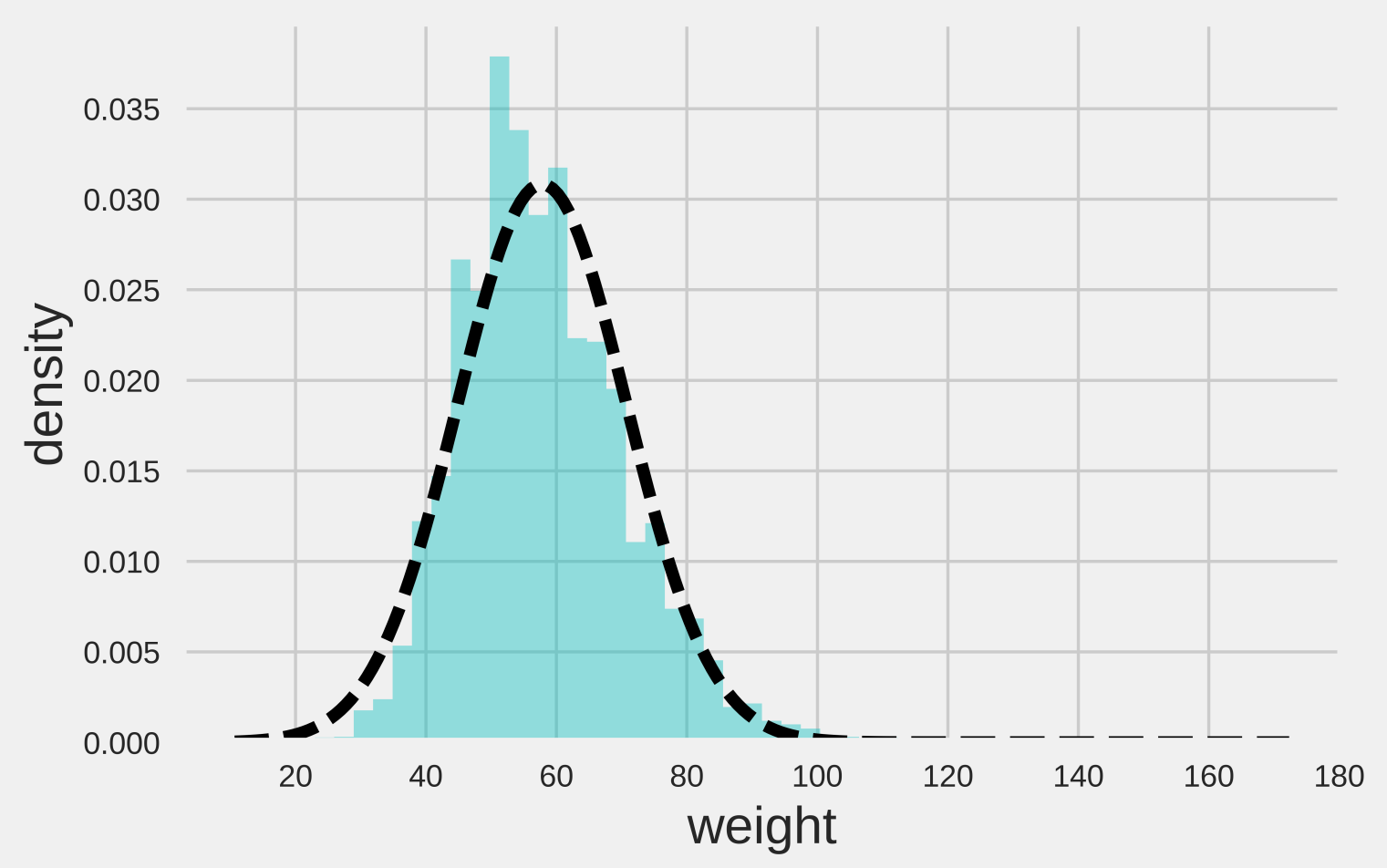
Histogram



Gambar: Contoh histogram dengan *kernel density estimation*

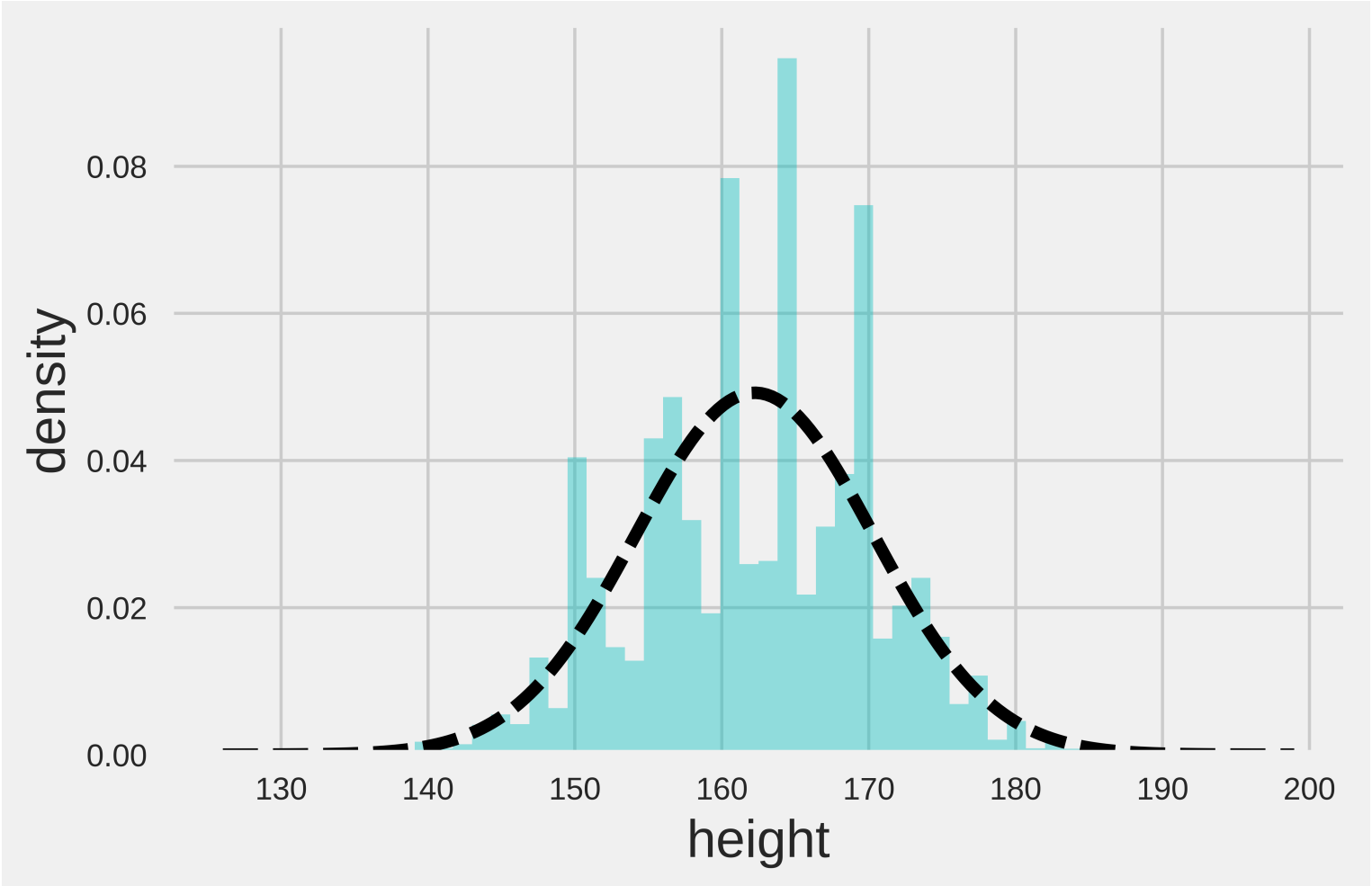
1. Digunakan untuk melihat distribusi dari variabel
2. Dibagi berdasarkan *bins*
3. Sangat bergantung pada jumlah *bins* yang digunakan!

Histogram



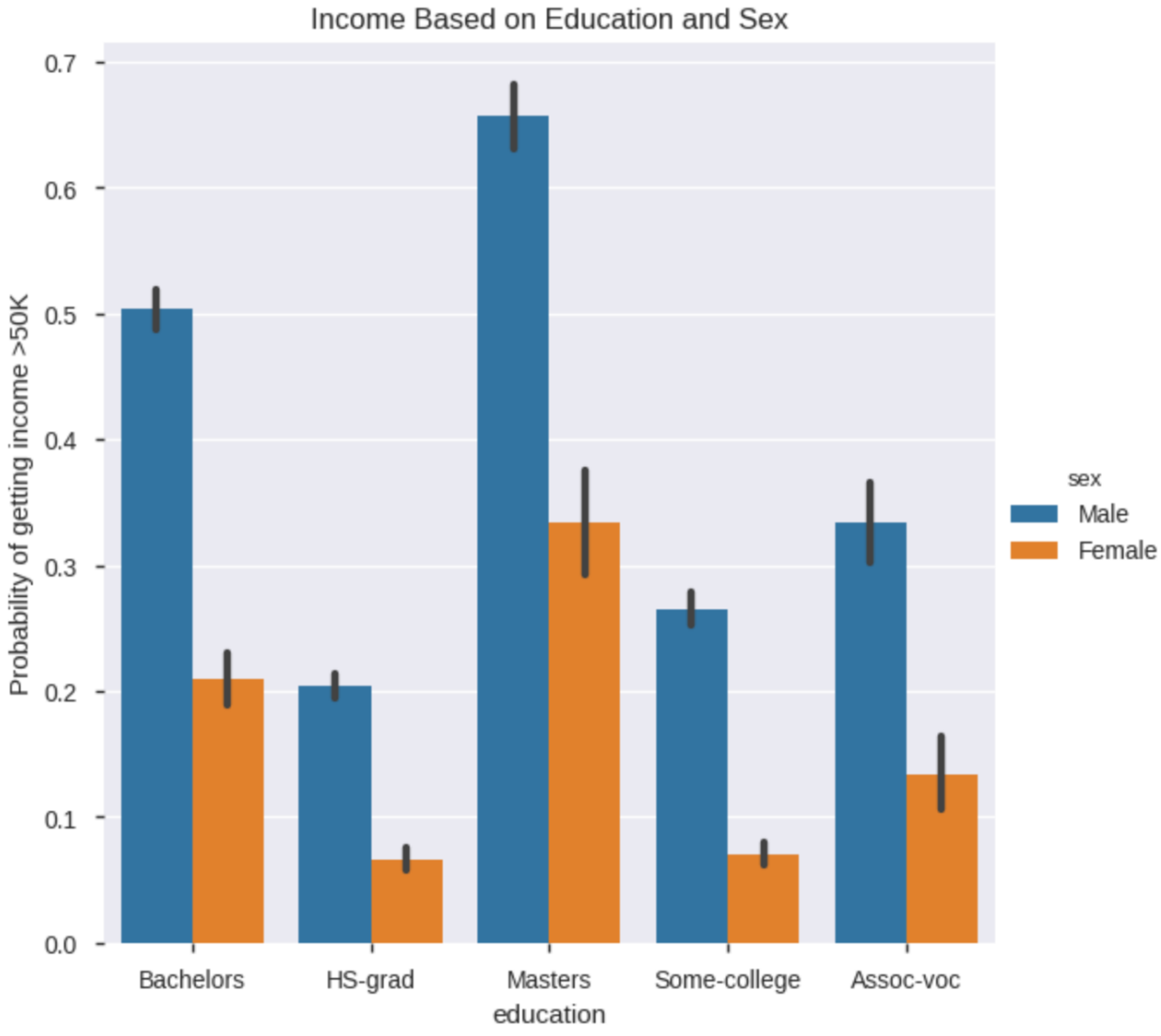
Gambar: Apa yang aneh dari histogram ini?

Histogram

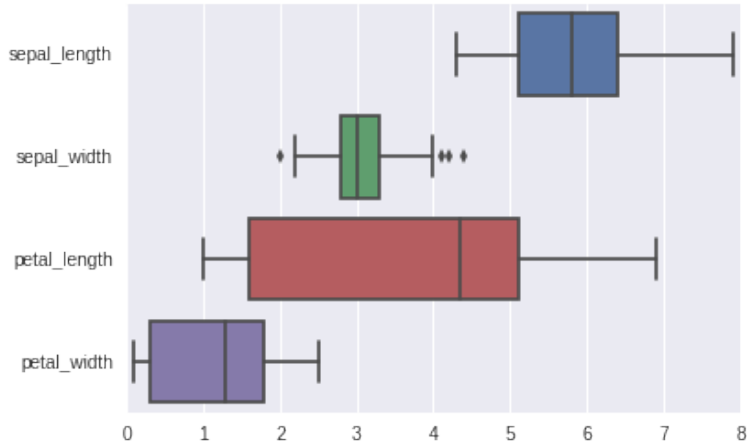


Gambar: Apa yang aneh dari histogram ini?

Bar Plot


Income Based on Education and Sex
Probability of getting income >50K
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Bachelors
HS-grad
Masters
Some-college
Assoc-voc
education
sex
Male
Female

Box Plot



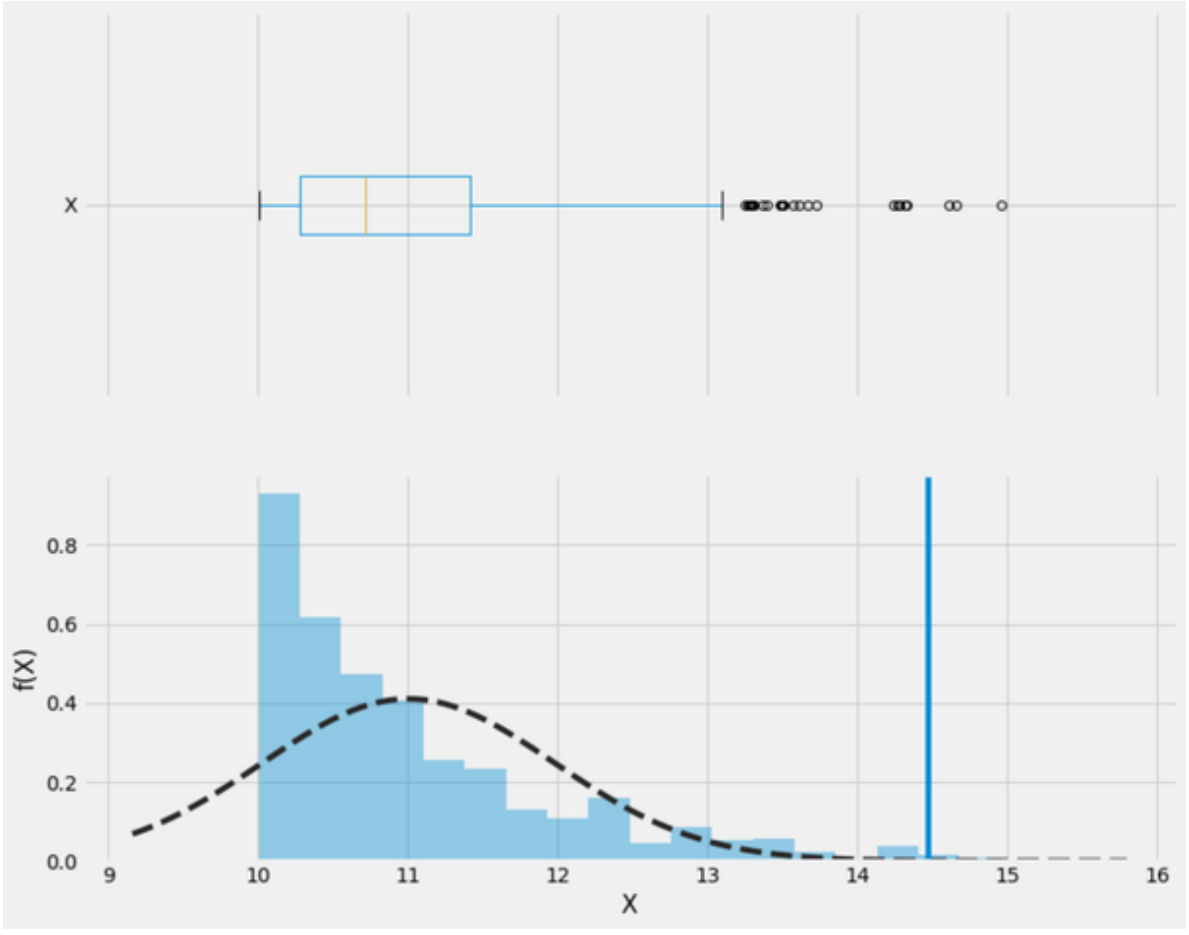
Gambar: Box plot untuk membandingkan atribut

1. Menggambarkan jangkauan dan persentil
2. Dapat digunakan untuk membandingkan atribut
3. Membantu menemukan pencilan

Potensi Masalah pada Box Plot

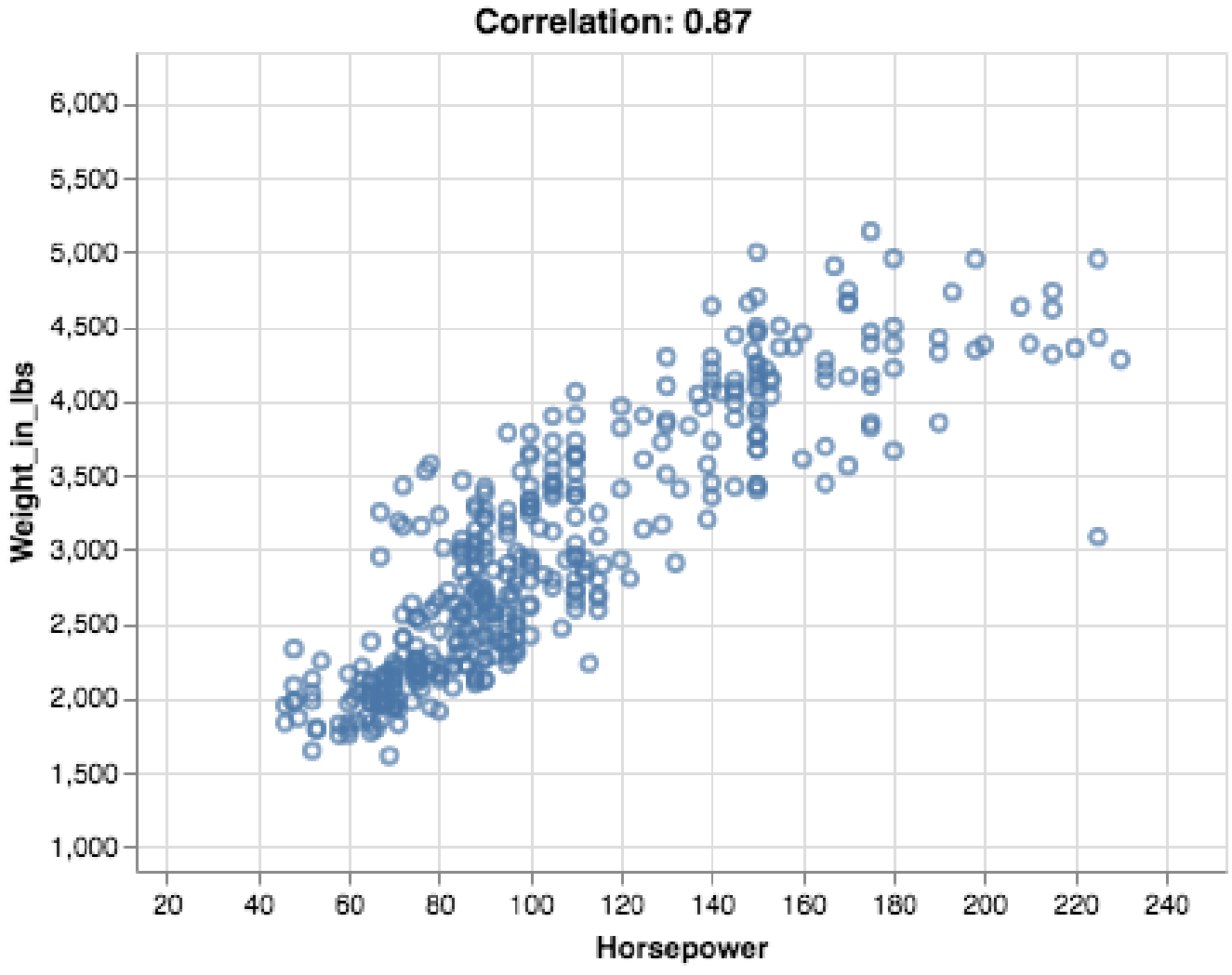
https://twitter.com/randal_olson/status/895678675814764544

Box Plot vs. Histogram

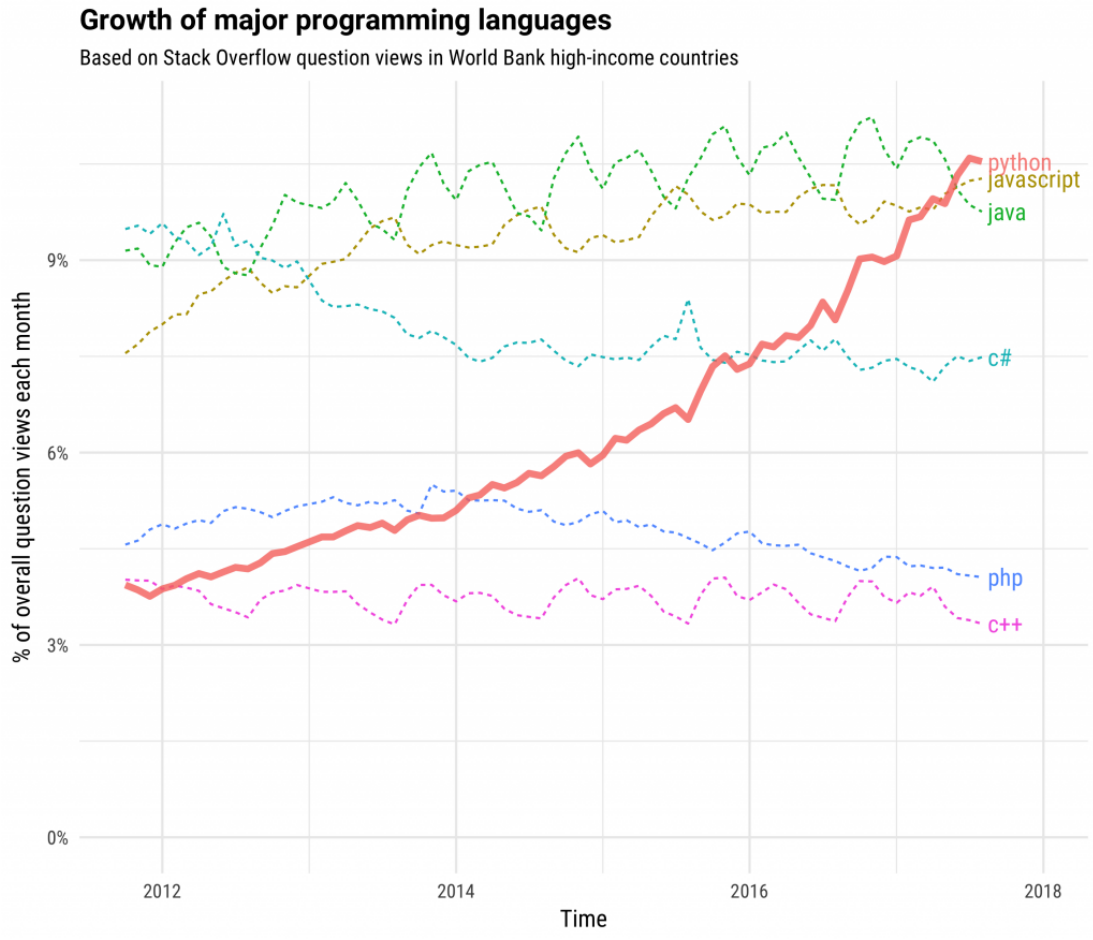


Gambar: Perbandingan box plot dan histogram untuk data yang sama

Scatter Plot

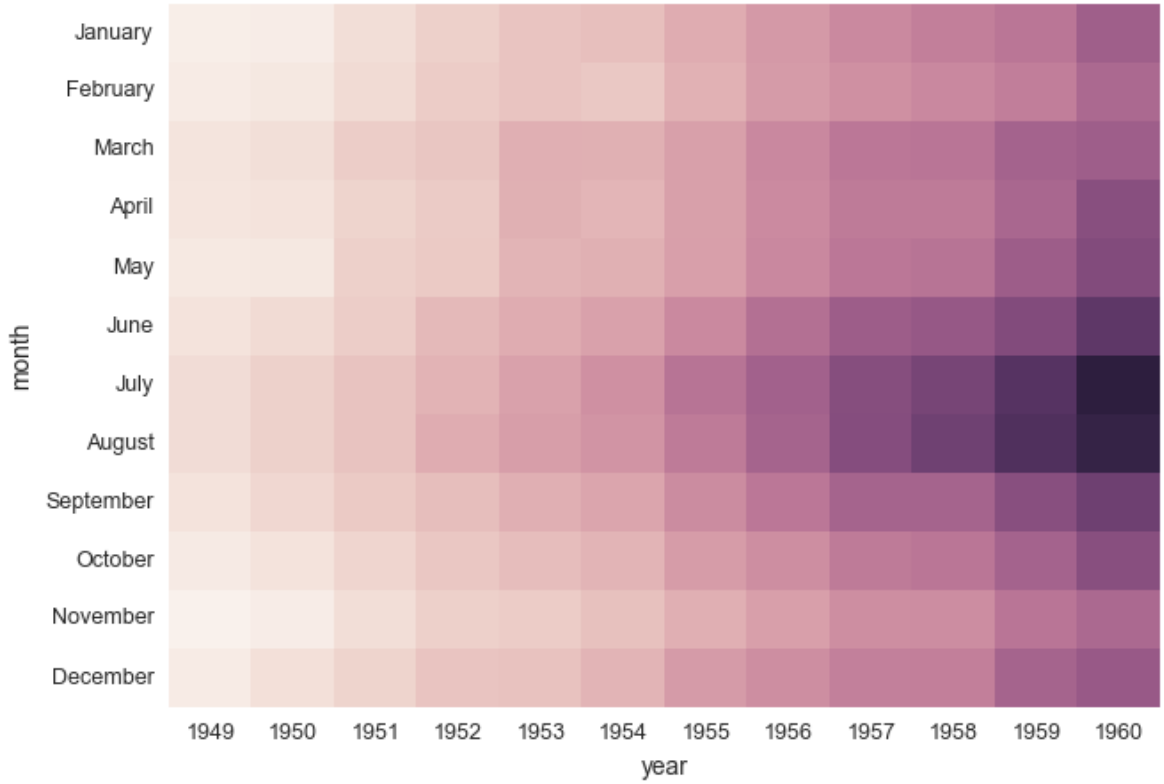


Timeseries



Gambar: Perubahan persentase pengunjung pertanyaan berdasarkan waktu [Robinson, 2017]

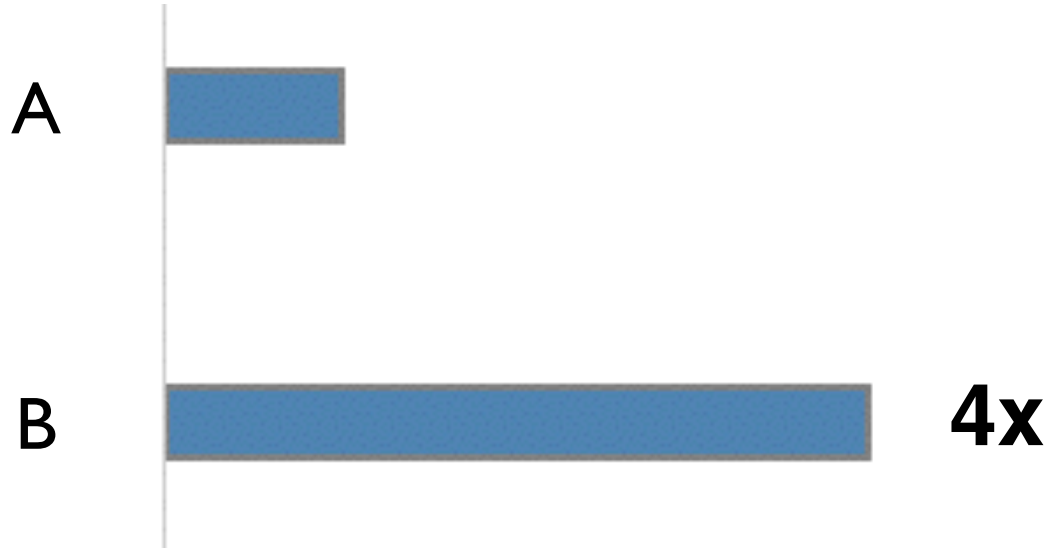
Heatmap



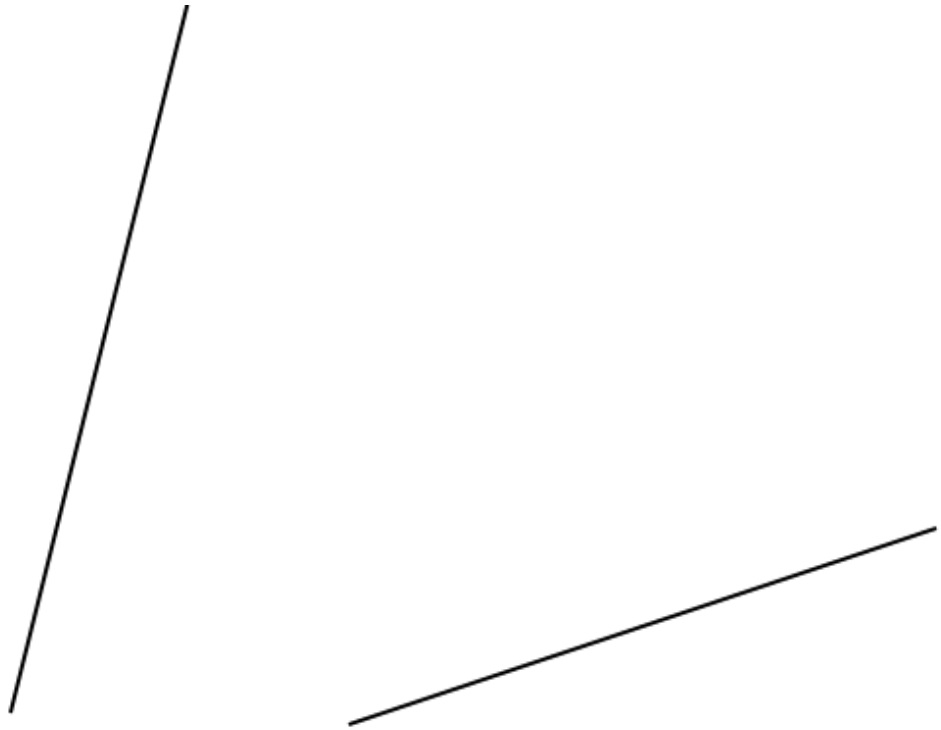
Gambar: Aktivitas penerbangan berdasarkan dua dimensi waktu

Visualisasi Efektif

How much longer?



How much steeper slope?

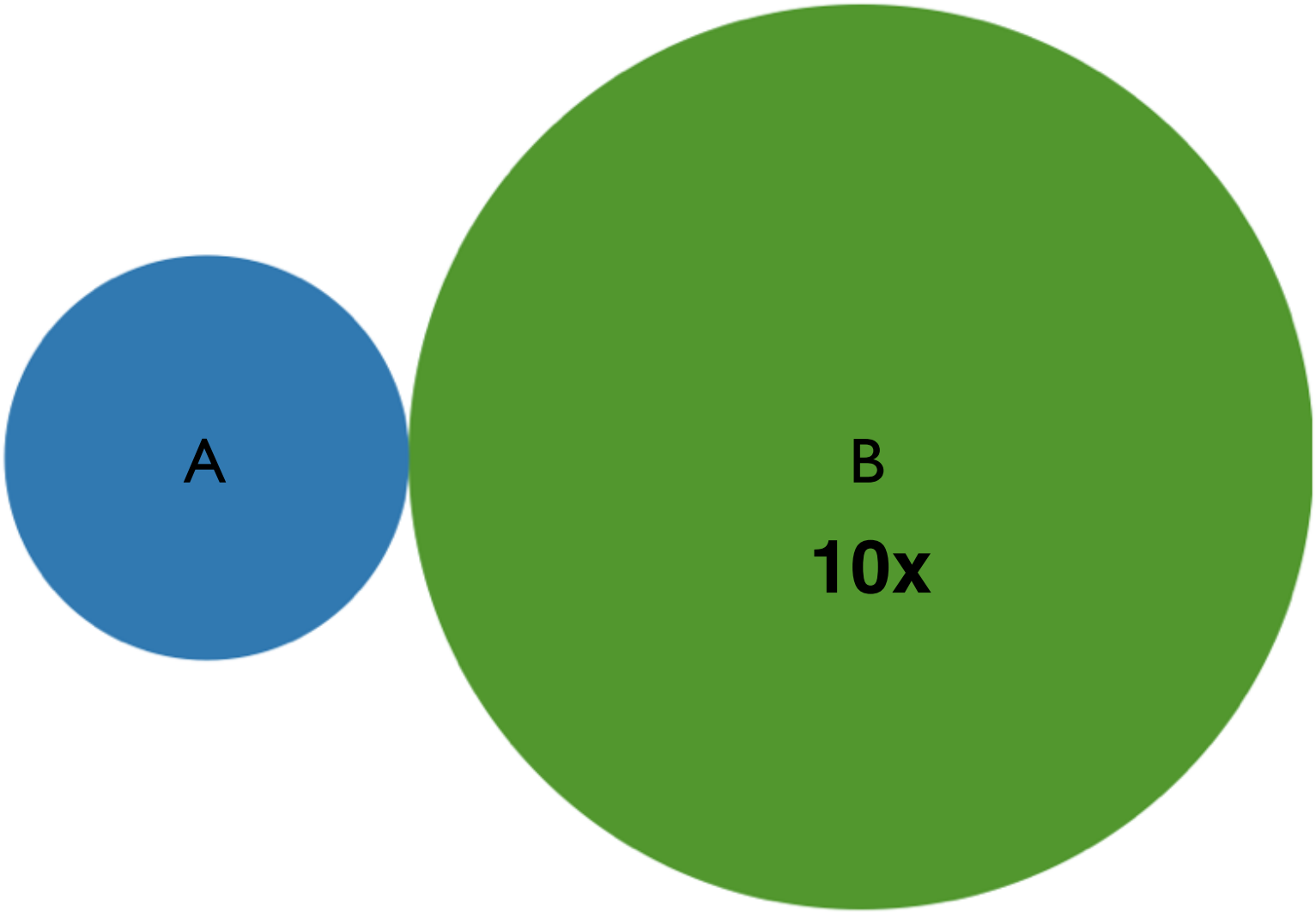


A

B

4x

How much larger area?



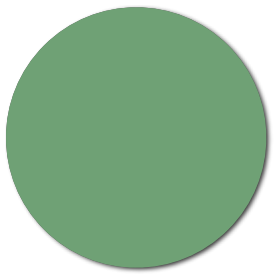
How much darker?

A

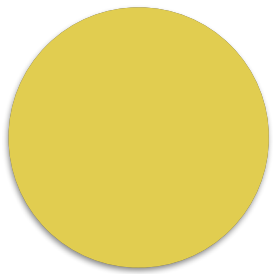
B

2x

How much bigger value?

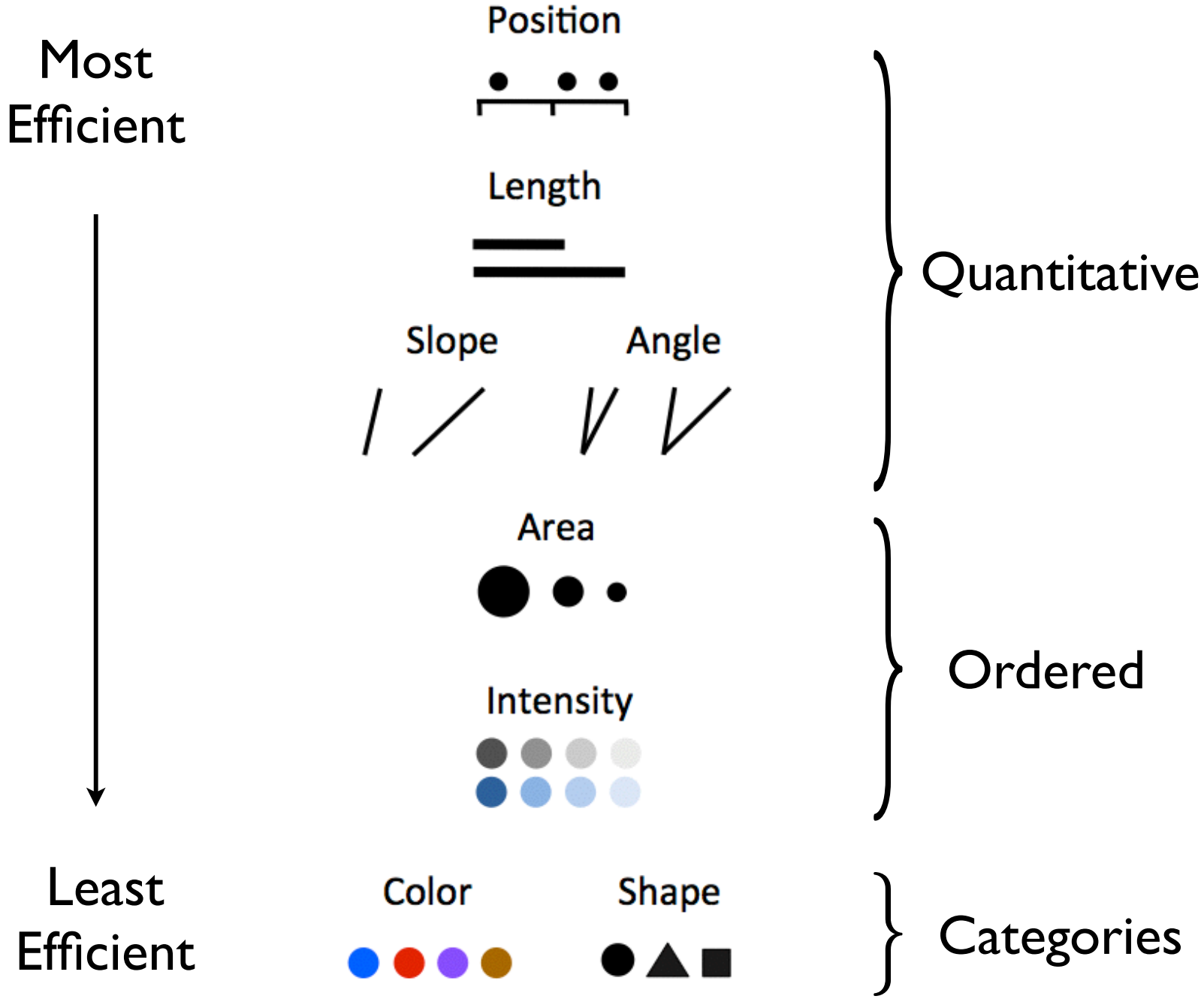


A

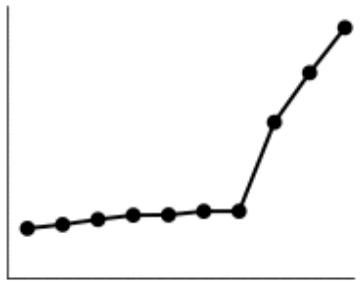


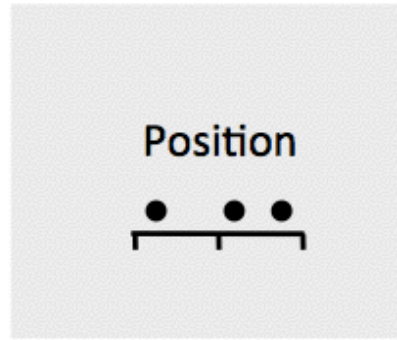
B

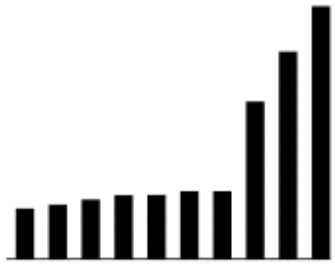
4x

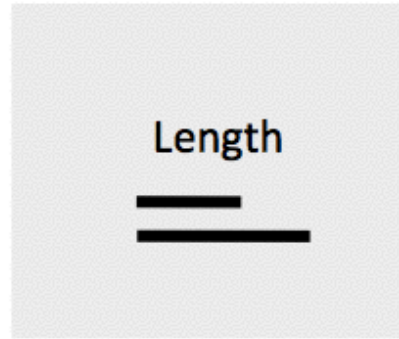

Most Efficient
Position
Length
Slope
Angle
Quantitative
Area
Intensity
Ordered
Least Efficient
Color
Shape
Categories

Most Effective

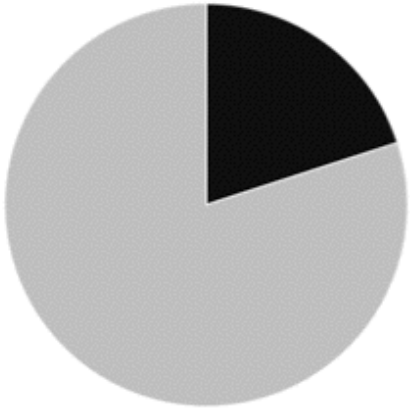


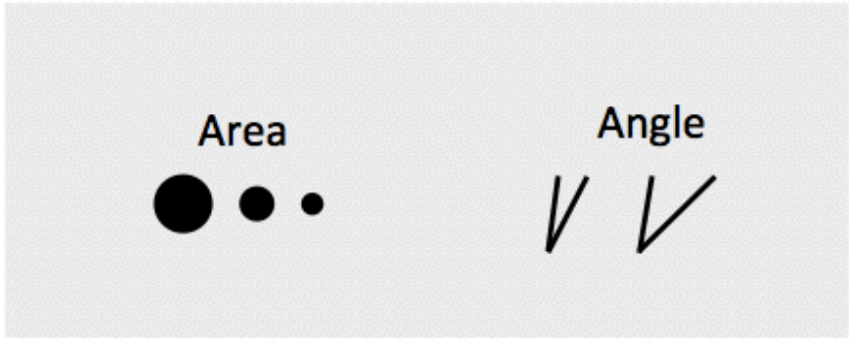


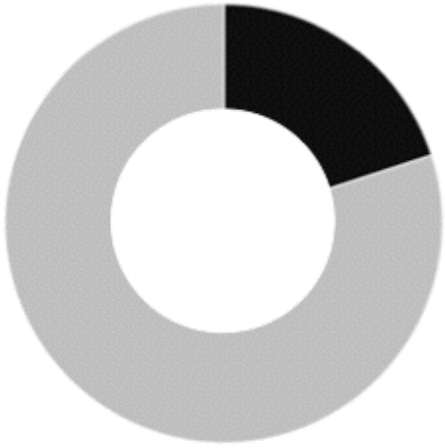


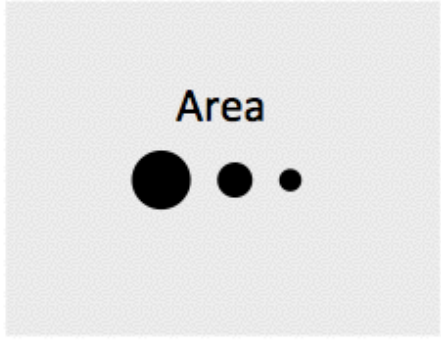


Less Effective



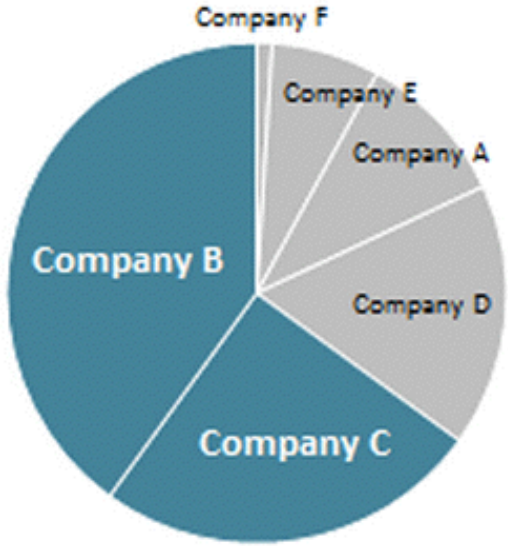


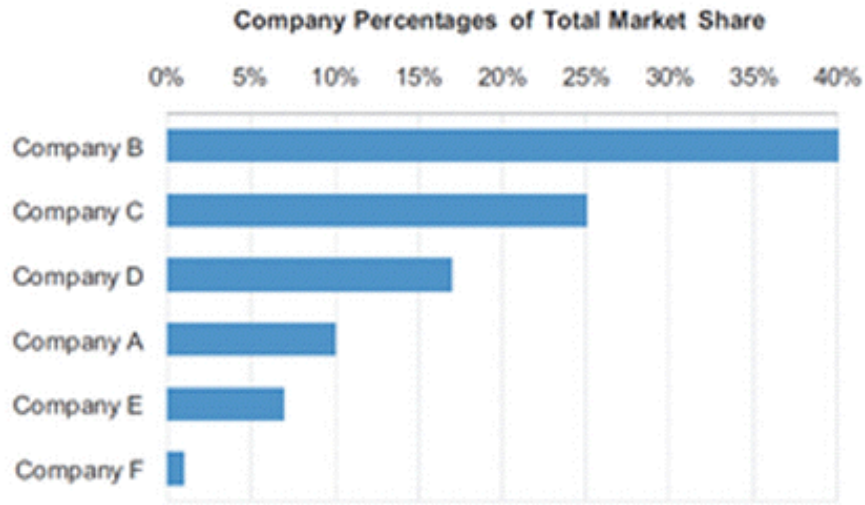




Pie vs. Bar Charts

65% of the market is controlled by companies B and C





Least Effective

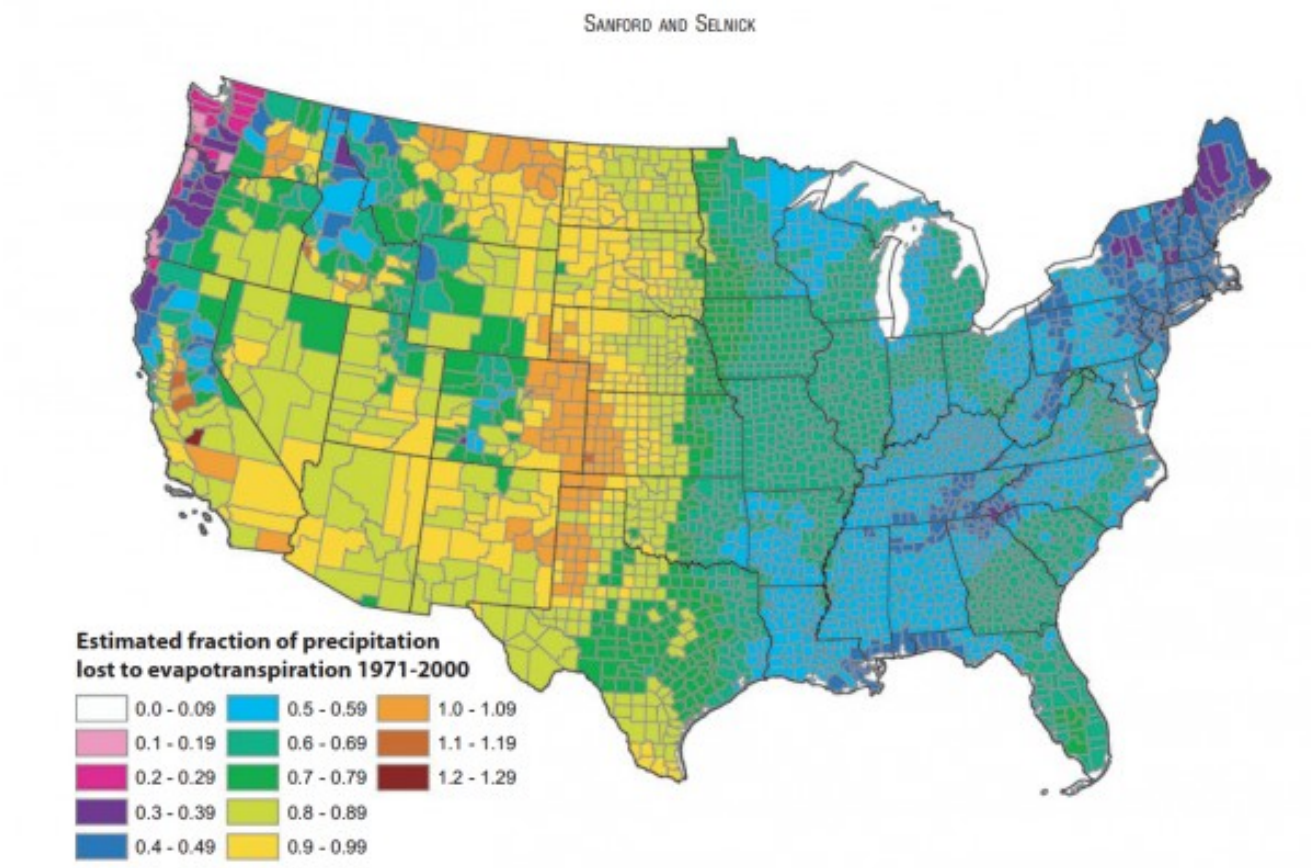


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

CS109 Data Science
Lecture 3: Exploratory Data Analysis

Referensi

Michael Waskom (2015)

Anscombe's quartet

http://seaborn.pydata.org/examples/anscombes_quartet.html

David Robinson (6 September 2017)

The Incredible Growth of Python

https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Terima kasih