Konsep Jarak dan Eksplorasi Data

Ali Akbar Septiandri

November 10, 2017

untuk Astra Graphia IT

Daftar Isi

Konsep Jarak Antardata

Mengapa kita perlu mengukur jarak antardata?

Mencari Data yang Mirip

1. Merupakan **permasalahan fundamental** untuk berbagai tugas dalam *data mining*, e.g. *clustering*, sistem rekomendasi, pengecekan plagiarisme

Mencari Data yang Mirip

1. Merupakan **permasalahan fundamental** untuk berbagai tugas dalam *data mining*, e.g. *clustering*, sistem rekomendasi, pengecekan plagiarisme
2. Kita ingin mengetahui **nilai** terkuantifikasi perbedaan atau kesamaan dari sepasang data

Mencari Data yang Mirip

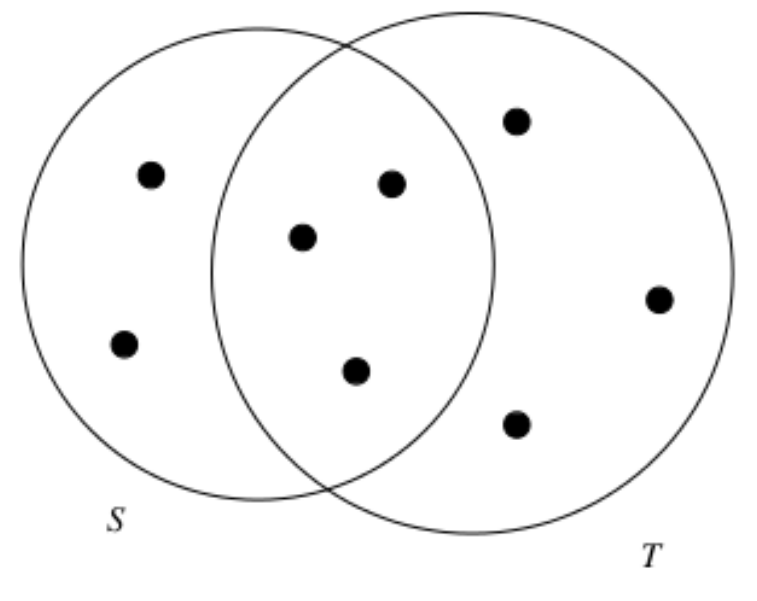
1. Merupakan **permasalahan fundamental** untuk berbagai tugas dalam *data mining*, e.g. *clustering*, sistem rekomendasi, pengecekan plagiarisme
2. Kita ingin mengetahui **nilai** terkuantifikasi perbedaan atau kesamaan dari sepasang data
3. Pengecekan untuk setiap pasang data bisa sangat merepotkan sehingga perlu **penyempitan pencarian**

Mencari Data yang Mirip

1. Merupakan **permasalahan fundamental** untuk berbagai tugas dalam *data mining*, e.g. *clustering*, sistem rekomendasi, pengecekan plagiarisme
2. Kita ingin mengetahui **nilai** terkuantifikasi perbedaan atau kesamaan dari sepasang data
3. Pengecekan untuk setiap pasang data bisa sangat merepotkan sehingga perlu **penyempitan pencarian**
4. Biasanya direpresentasikan dalam nilai $[0, 1]$

Jaccard Similarity

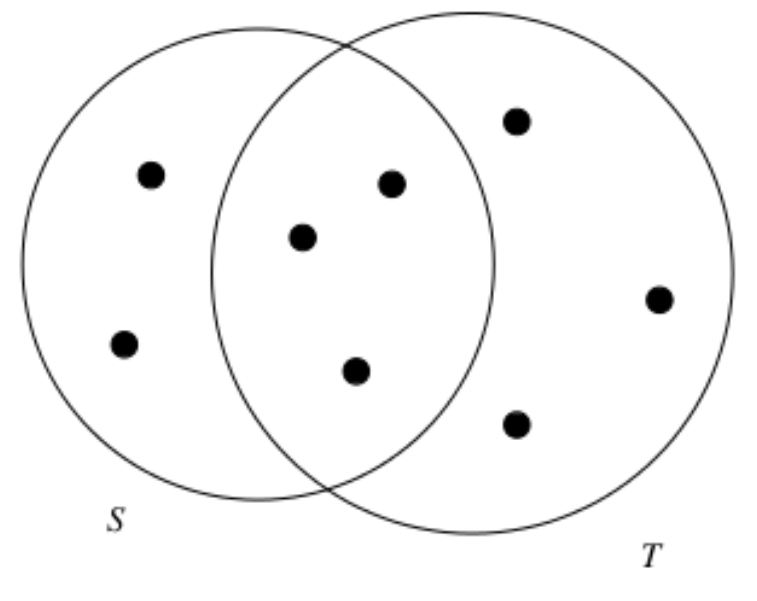
1. Jaccard similarity dari himpunan S dan T adalah $SIM(S, T) = |S \cap T|/|S \cup T|$



Gambar 1: Dua himpunan dengan *Jaccard similarity* 3/8 [Leskovec, et al. 2014]

Jaccard Similarity

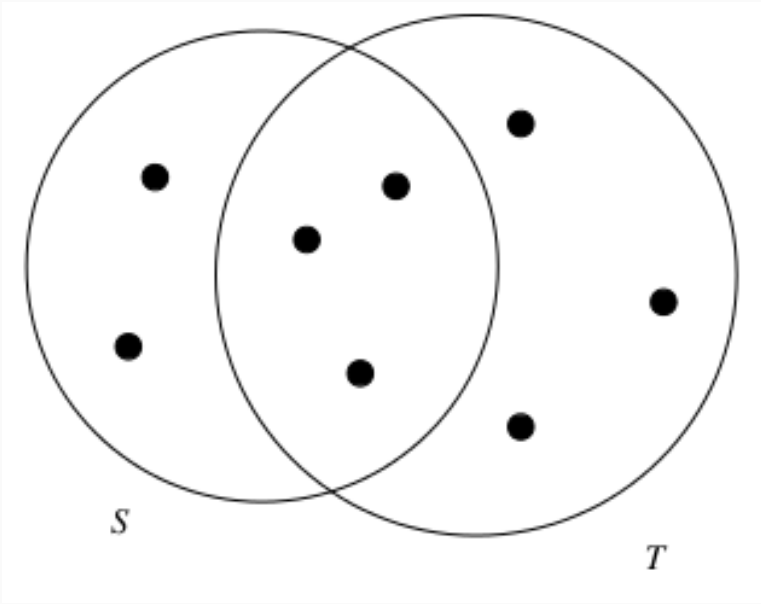
1. Jaccard similarity dari himpunan S dan T adalah $SIM(S, T) = |S \cap T|/|S \cup T|$
2. Dapat digunakan untuk menemukan sepasang dokumen yang **mirip secara leksikal**



Gambar 1: Dua himpunan dengan *Jaccard similarity* 3/8 [Leskovec, et al. 2014]

Jaccard Similarity

1. Jaccard similarity dari himpunan S dan T adalah $SIM(S, T) = |S \cap T|/|S \cup T|$
2. Dapat digunakan untuk menemukan sepasang dokumen yang **mirip secara leksikal**
3. Berguna juga dalam *collaborative filtering*



Gambar 1: Dua himpunan dengan *Jaccard similarity* 3/8 [Leskovec, et al. 2014]

Jaccard Similarity dari Dua Vektor

1. Berhati-hatilah saat membandingkan dua vektor biner dengan Jaccard similarity!
2. Akan banyak kesamaan nilai 0 yang ditemukan
3. *Jaccard similarity* $\neq$ *simple matching*

Collaborative Filtering dengan Kemiripan Himpunan

1. Dalam kasus belanja *online*, jarang ditemukan dua orang dengan Jaccard similarity yang besar

Collaborative Filtering dengan Kemiripan Himpunan

1. Dalam kasus belanja *online*, jarang ditemukan dua orang dengan Jaccard similarity yang besar
2. Nilai 20% pada Jaccard similarity antara dua orang sudah bisa dianggap signifikan [Leskovec, et al. 2014]

Collaborative Filtering dengan Kemiripan Himpunan

1. Dalam kasus belanja *online*, jarang ditemukan dua orang dengan Jaccard similarity yang besar
2. Nilai 20% pada Jaccard similarity antara dua orang sudah bisa dianggap signifikan [Leskovec, et al. 2014]
3. Perlu penyesuaian jika datanya didasarkan dari *rating*

Collaborative Filtering pada Kasus Peringkat Film

Beberapa opsi yang bisa dipilih saat merepresentasikan nilai atribut saat didasarkan pada pemberian peringkat film
[Leskovec, et al. 2014]:

1. Membuang film yang diberi peringkat rendah - anggap tidak pernah ditonton

Collaborative Filtering pada Kasus Peringkat Film

Beberapa opsi yang bisa dipilih saat merepresentasikan nilai atribut saat didasarkan pada pemberian peringkat film
[Leskovec, et al. 2014]:

1. Membuang film yang diberi peringkat rendah - anggap tidak pernah ditonton
2. Menggunakan dua himpunan per film: “suka” dan “tidak suka”

Collaborative Filtering pada Kasus Peringkat Film

Beberapa opsi yang bisa dipilih saat merepresentasikan nilai atribut saat didasarkan pada pemberian peringkat film
[Leskovec, et al. 2014]:

1. Membuang film yang diberi peringkat rendah - anggap tidak pernah ditonton
2. Menggunakan dua himpunan per film: "suka" dan "tidak suka"
3. Jika menggunakan sistem lima bintang, masukkan film ke dalam himpunan seorang pengguna n kali jika film tersebut diberikan n bintang*

***Poin terakhir menyebabkan perhitungannya harus menggunakan Jaccard similarity for bags**

Jaccard Similarity for Bags

Example ([Leskovec, et al. 2014])

Bag-similarity dari *bags* $\{a, a, a, b\}$ dan $\{a, a, b, b, c\}$ adalah 1/3. Irisannya akan mencacah **dua kemunculan** a dan **satu kemunculan** b, i.e. 3. Gabungannya adalah jumlah total elemen kedua *bags*, i.e. 9.

Jaccard Similarity for Bags

Example ([Leskovec, et al. 2014])

Bag-similarity dari *bags* $\{a, a, a, b\}$ dan $\{a, a, b, b, c\}$ adalah 1/3. Irisannya akan mencacah **dua kemunculan** a dan **satu kemunculan** b, i.e. 3. Gabungannya adalah jumlah total elemen kedua *bags*, i.e. 9.

Pertanyaan

Berapa nilai maksimal dari dua *bags* yang sama?

Cosine Similarity

Definisi

Jika d_1 dan d_2 adalah vektor dokumen, maka
$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$ dengan $\|d\|$ adalah panjang vektor d.

Properti

Nilai dari cosine similarity:

1. 1 - kedua vektor sama
2. 0 - kedua vektor tegak lurus
3. -1 - kedua vektor bertolak belakang

Cosine Similarity

Definisi

Jika d_1 dan d_2 adalah vektor dokumen, maka
$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$ dengan $\|d\|$ adalah panjang vektor d.

Properti

Nilai dari cosine similarity:

1. 1 - kedua vektor sama
2. 0 - kedua vektor tegak lurus
3. -1 - kedua vektor bertolak belakang

Pertanyaan

Kapan cosine similarity lebih dipilih dibandingkan Jaccard similarity?

Jarak

Pengukuran jarak (*distance measures*) didefinisikan sebagai fungsi $d(x, y)$ yang menerima dua titik sebagai argumen dan mengembalikan nilai riil. Beberapa properti yang dimiliki jarak antara lain:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0$ jika dan hanya jika $x = y$
3. $d(x, y) = d(y, x)$ (simetris)
4. $d(x, y) \leq d(x, z) + d(z, y)$ (ketaksamaan segitiga)

Euclidean Distance

Definisi

$$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

yang sering juga dirujuk sebagai L_2-norm

Euclidean Distance

Definisi

$$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

yang sering juga dirujuk sebagai L_2-norm

Pertanyaan

Kapan kita harus menggunakan Euclidean distance, kapan kita harus menggunakan cosine similarity?

Manhattan Distance

Gambar 2: Manhattan vs. Euclidean distance [Grigorev, 2015]

Manhattan Distance

Definisi

$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sum_{i=1}^{n} |x_i - y_i|$ (L_1-norm)

Minkowski Distance

Definisi

Dari dua bentuk tersebut, kita bisa melihat generalisasi rumusnya (L_r-norm) sebagai:

$$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sqrt[r]{\sum_{i=1}^{n} |x_i - y_i|^r}$$

Minkowski Distance

Definisi

Dari dua bentuk tersebut, kita bisa melihat generalisasi rumusnya (L_r-norm) sebagai:

$$d([x_1, x_2, ..., x_n], [y_1, y_2, ..., y_n]) = \sqrt[r]{\sum_{i=1}^{n} |x_i - y_i|^r}$$

Pertanyaan

Apa yang terjadi saat $r \to \infty$?

Jaccard dan Cosine Distance

Similarity & Distance

Similarities pada dasarnya dapat diubah menjadi *distances*

Jaccard dan Cosine Distance

Similarity & Distance

Similarities pada dasarnya dapat diubah menjadi *distances*

Jaccard distance

$$d(S, T) = 1 - SIM(S, T)$$

Jaccard dan Cosine Distance

Similarity & Distance

Similarities pada dasarnya dapat diubah menjadi *distances*

Jaccard distance

$$d(S, T) = 1 - SIM(S, T)$$

Cosine distance

$$cos(\theta) = cos(d_1, d_2) = y;$$
$$\theta = cos^{-1}(y)$$

Mahalanobis Distance

Definisi

Mahalanobis distance adalah jarak antara titik P dengan distribusi D (dengan rata-ratanya)

Formula

Untuk suatu titik $\vec{x} = (x_1, x_2, ..., x_N)^T$ dari suatu distribusi dengan rata-rata $\vec{\mu} = (\mu_1, \mu_2, ..., \mu_N)^T$ dengan matriks kovarian Σ didefinisikan sebagai:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$$

Generalisasi Mahalanobis Distance

Jarak antartitik

Untuk dua vektor acak $\vec{x}$ dan $\vec{y}$ yang berasal dari satu distribusi dengan matriks kovarian Σ:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$$

Generalisasi Mahalanobis Distance

Jarak antartitik

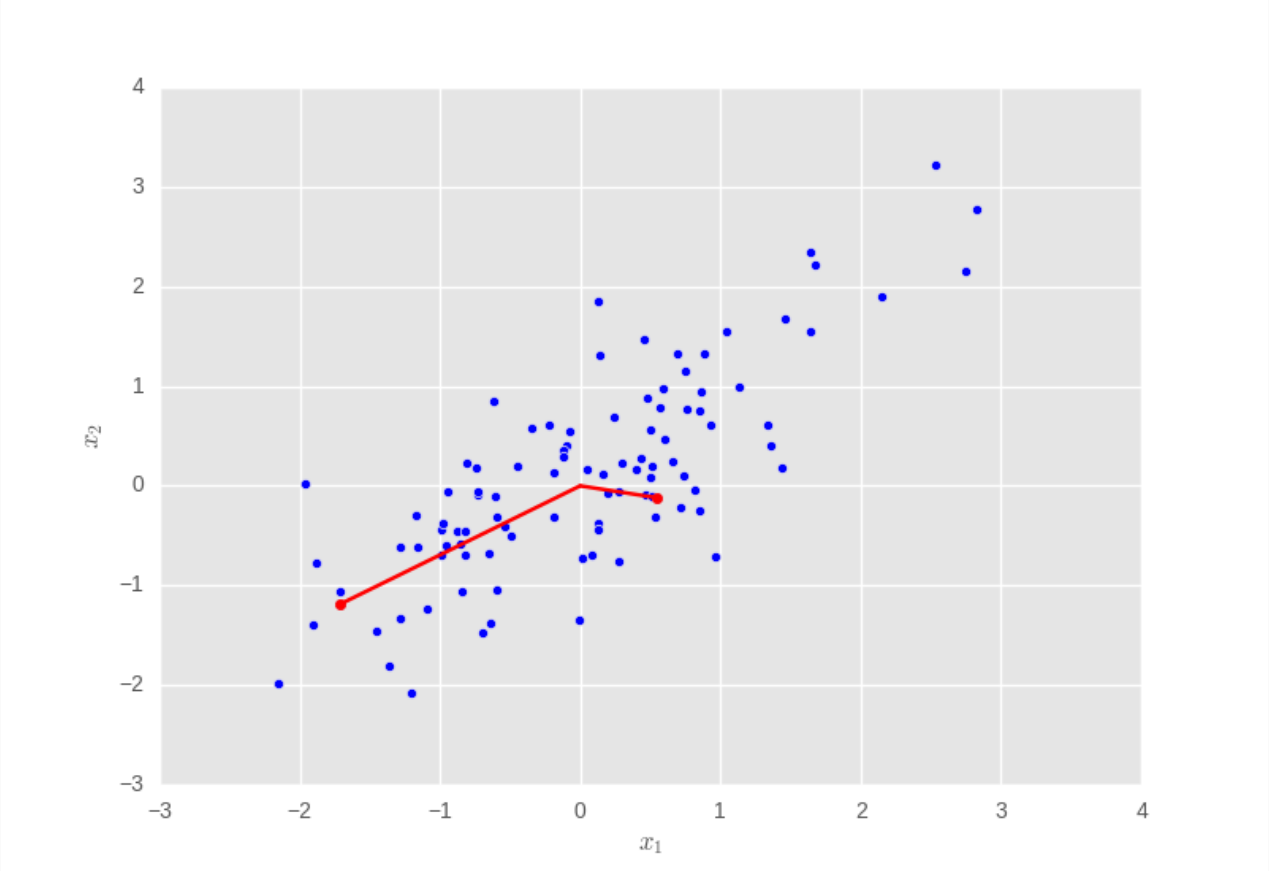
Untuk dua vektor acak $\vec{x}$ dan $\vec{y}$ yang berasal dari satu distribusi dengan matriks kovarian Σ:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$$

Euclidean distance

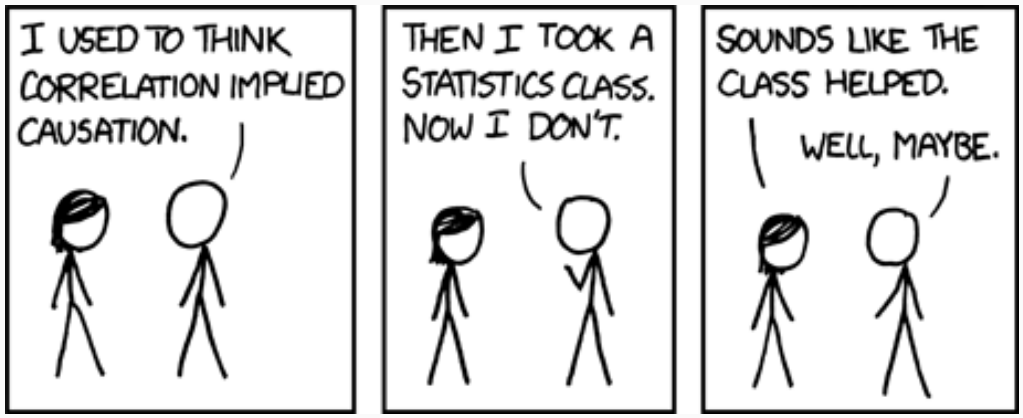
Perhatikan bahwa saat matriks kovariannya merupakan matriks identitas (I), maka Mahalanobis distance berubah menjadi Euclidean distance!

Mahalanobis Distance



Gambar 3: Mahalanobis distance antara dua titik dari Gaussian 2 dimensi

Korelasi



Gambar 4: Korelasi | Sumber: `https://xkcd.com/552/`

Korelasi

- Korelasi mengukur hubungan linear antarvariabel
- Dihitung dengan standardisasi data, p dan q, lalu menghitung produk skalarnya:

$$p'_i = \frac{p_i - \bar{p}}{std(p)}$$

$$q'_i = \frac{q_i - \bar{q}}{std(q)}$$

$$Cor(p, q) = p' \cdot q'$$

- Nilai korelasi ada di rentang [-1, 1]

Berhati-hatilah pada *spurious correlations*!

`http://www.tylervigen.com/spurious-correlations`

Eksplorasi Data

Frekuensi dan Modus

1. Perhitungan ini sering dilakukan pada data nominal
2. Modus adalah nilai yang paling sering muncul
3. Tidak memedulikan urutan data

Beberapa Nilai yang Penting

1. mean: $mean(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. median:
$$median(x) = \begin{cases} x_{r+1} & n \ mod \ 2 = 1, i.e. \ n = 2r+1 \\ \frac{1}{2}(x_r + x_{r+1}) & n \ mod \ 2 = 0, i.e. \ n = 2r \end{cases}$$
3. jangkauan: $range(x) = max(x) - min(x)$
4. varians: $var(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Visualisasikan!

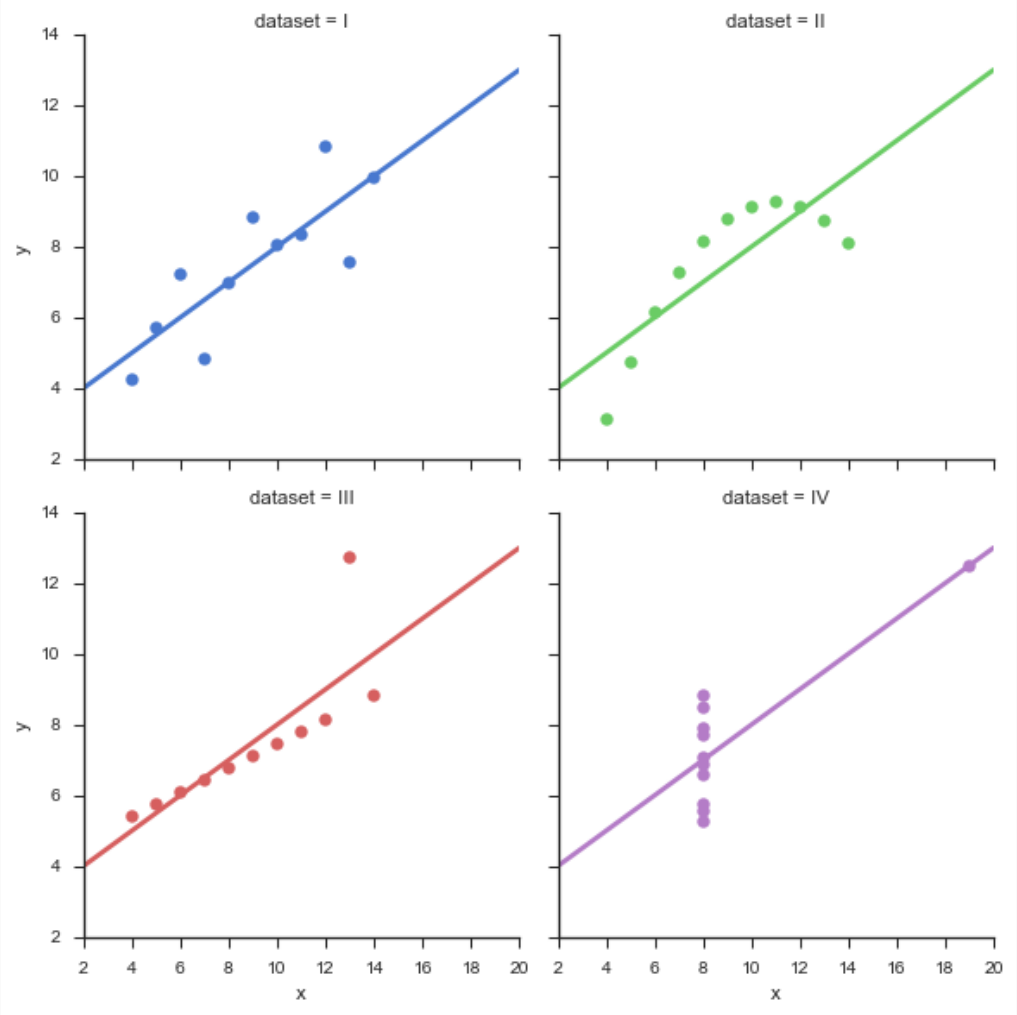
Pentingnya Visualisasi

Visualisasi dapat membantu:

1. mendeteksi pola dan tren secara umum
2. menemukan pencilan dan anomali
3. sangat mudah bagi manusia saat terlihat secara visual!

Expect problems in your data!

Visualisasi vs Summary Statistics



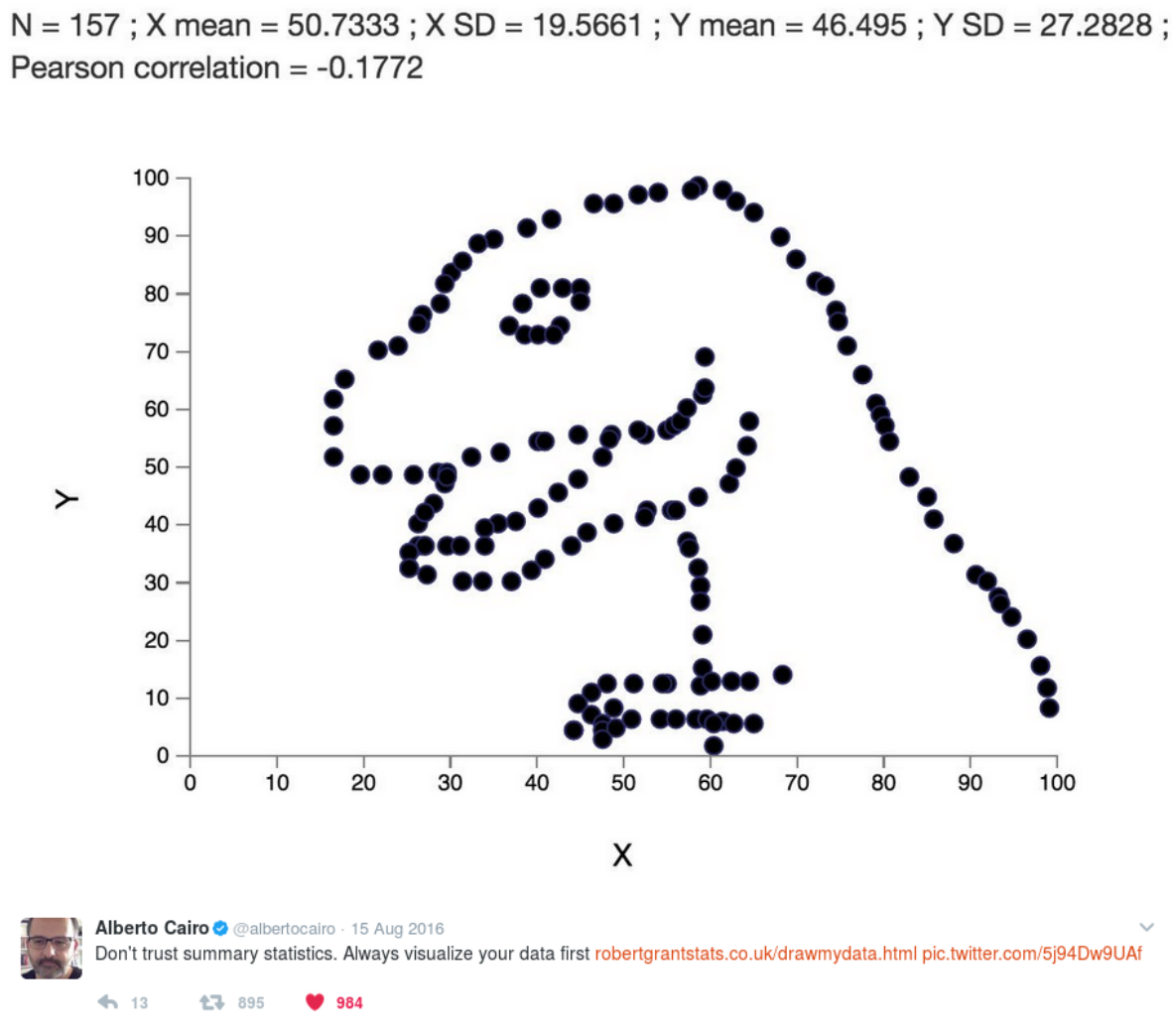
Gambar 5: *The infamous Anscombe's quartet* [Waskom, 2015]

Keempat data tersebut mempunyai
rataan, simpangan baku, dan nilai korelasi yang sama!

Don't trust summary statistics.

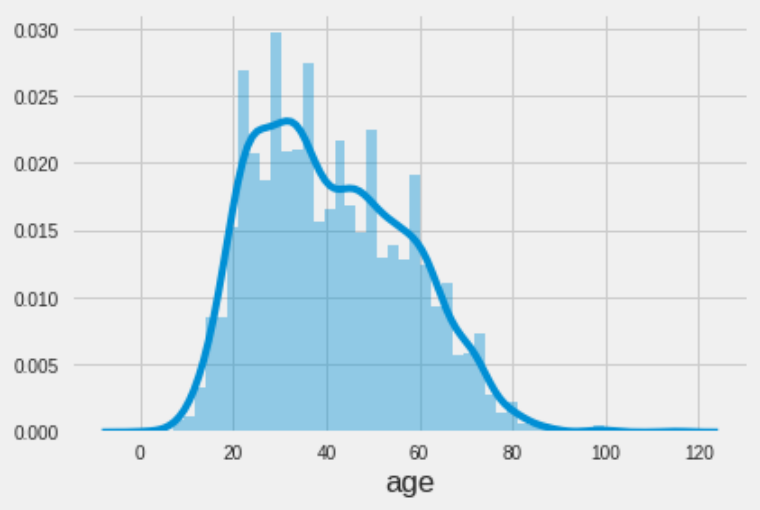
Always visualize your data first!

Visualisasi vs Summary Statistics



Gambar 6: Pentingnya visualisasi

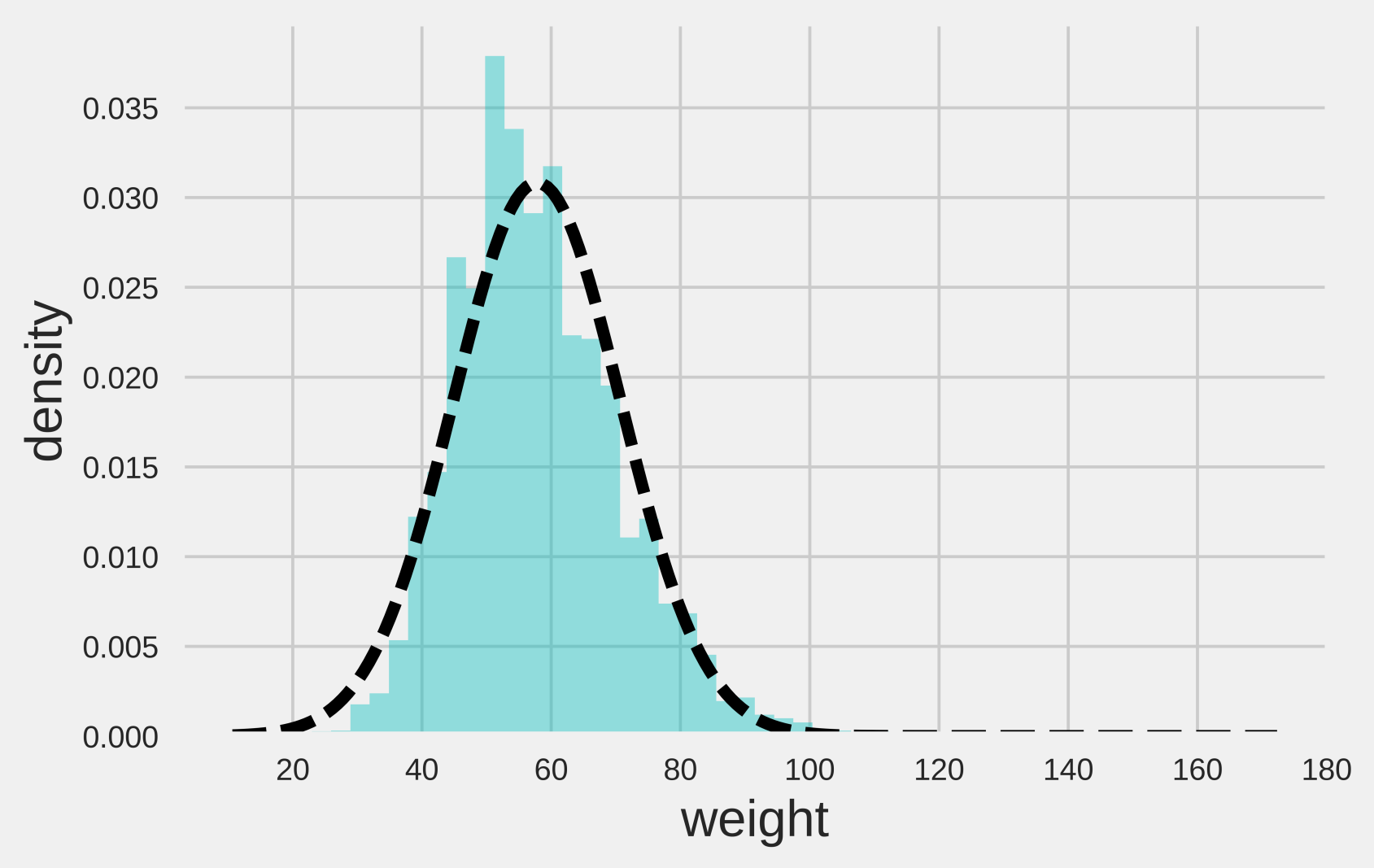
Histogram



Gambar 7: Contoh histogram dengan *kernel density estimation*

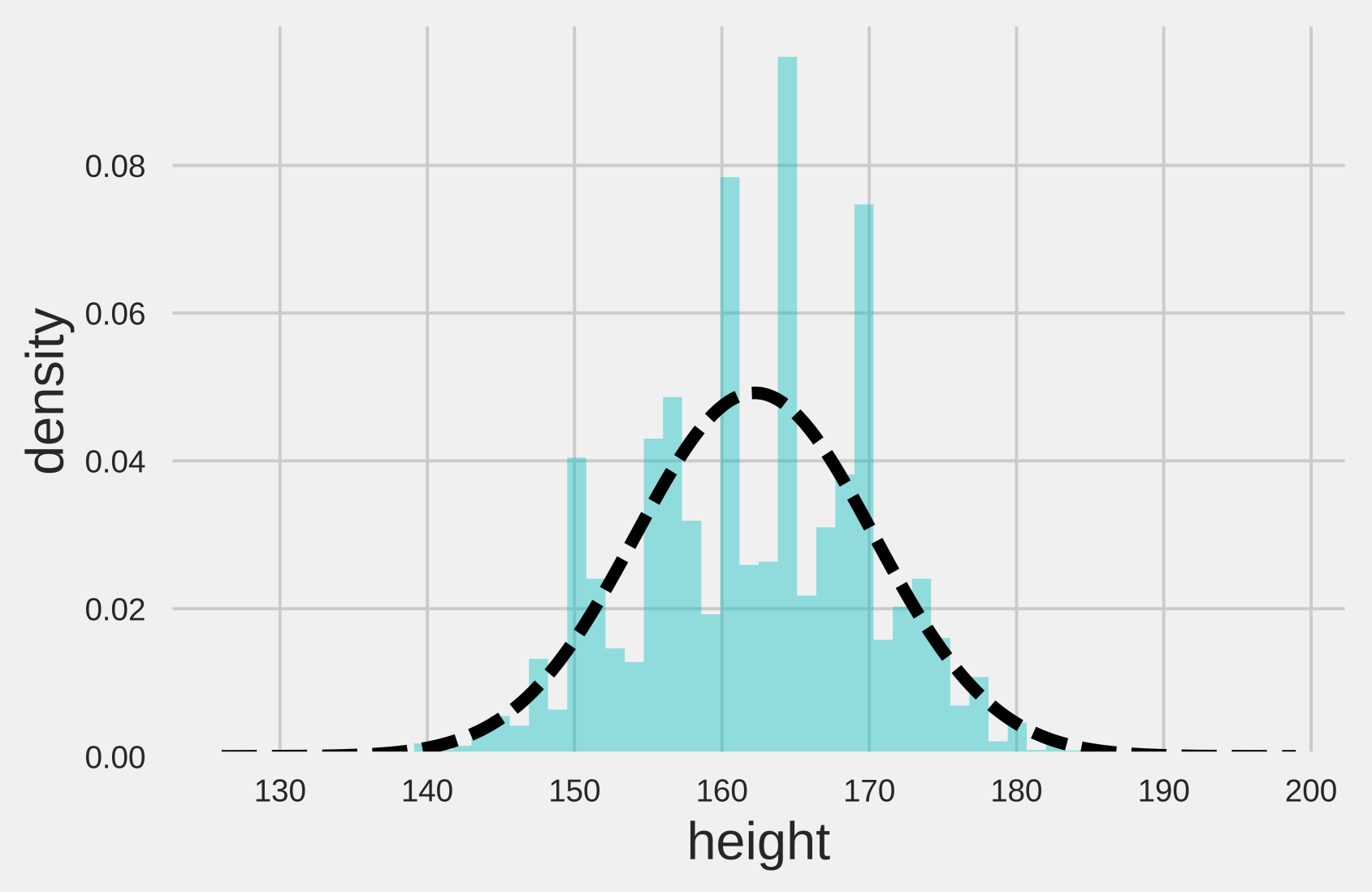
1. Digunakan untuk melihat distribusi dari variabel
2. Dibagi berdasarkan *bins*
3. Sangat bergantung pada jumlah *bins* yang digunakan!

Histogram



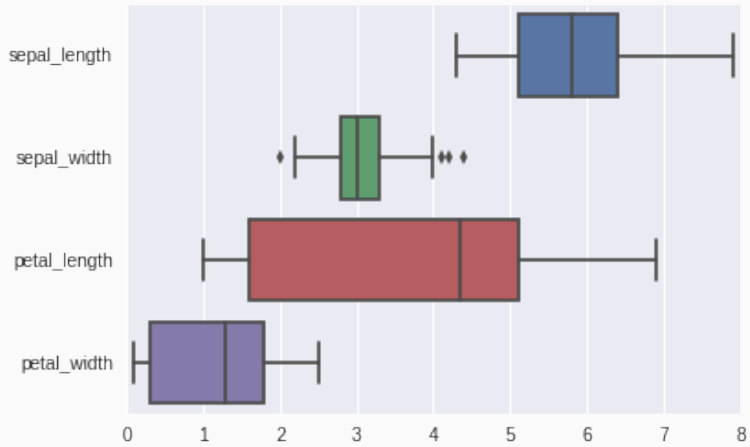
Gambar 8: Apa yang aneh dari histogram ini?

Histogram



Gambar 9: Apa yang aneh dari histogram ini?

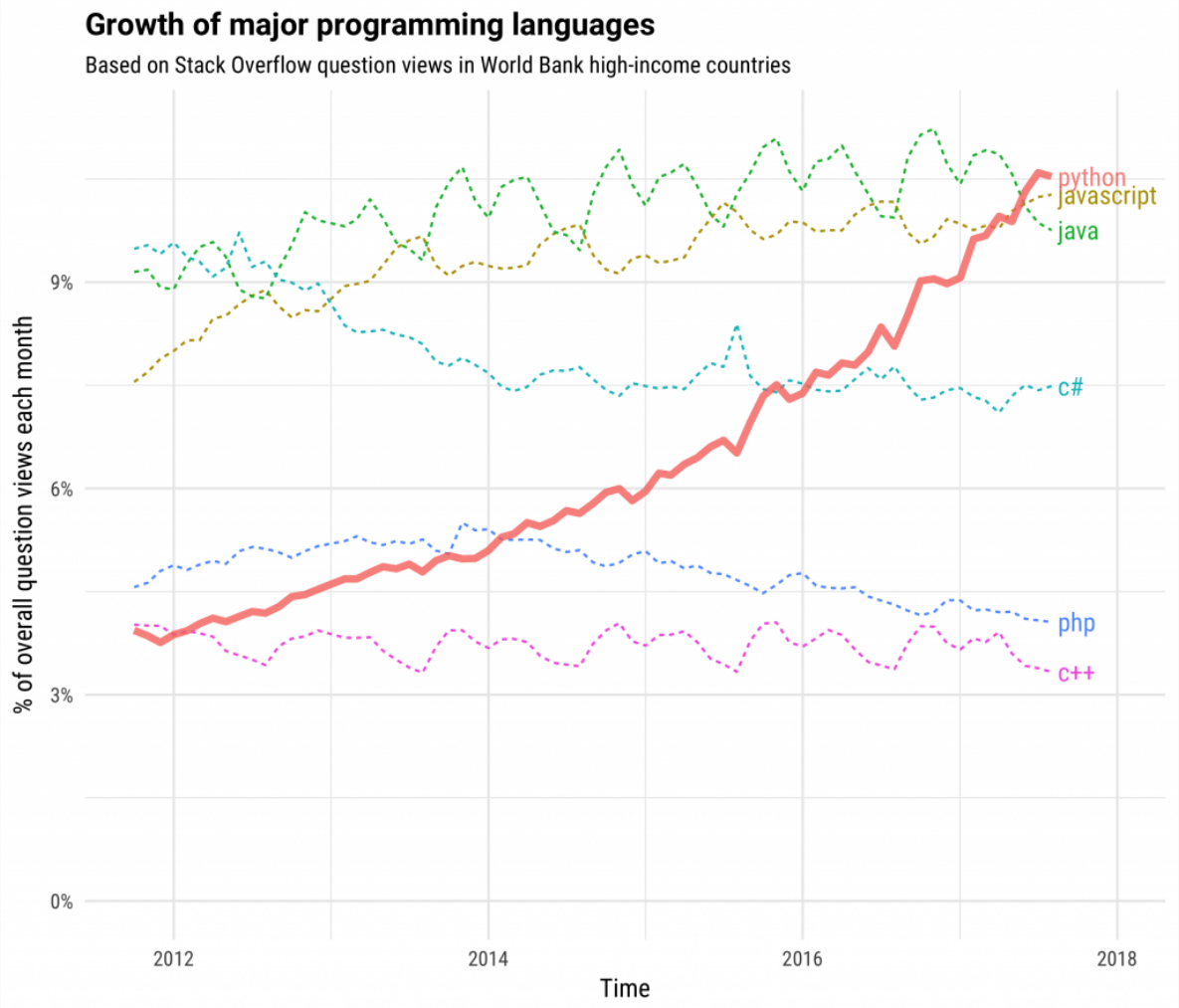
Box Plot



Gambar 10: Box plot untuk membandingkan atribut

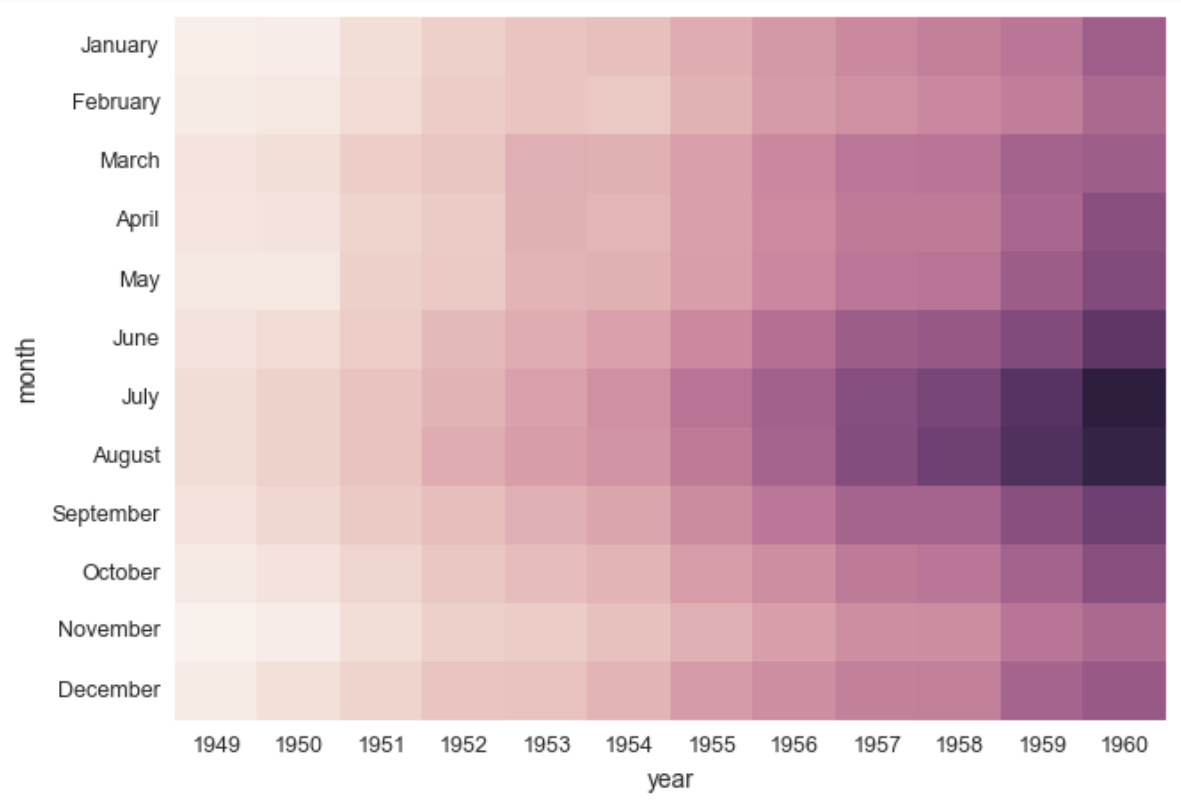
1. Menggambarkan jangkauan dan persentil
2. Dapat digunakan untuk membandingkan atribut
3. Membantu menemukan pencilan

Timeseries



Gambar 11: Perubahan persentase pengunjung pertanyaan berdasarkan waktu [Robinson, 2017]

Heatmap



Gambar 12: Aktivitas penerbangan berdasarkan dua dimensi waktu

Praktikum

Deskripsi Dataset

- Pokemon dataset
- Pembuat: Alberto Barradas (2016)
- `https://www.kaggle.com/abcsds/pokemon`
- Atribut: nama, tipe 1, tipe 2, HP, attack, defense, sp atk, sp def, speed, generasi
- Kandidat kelas: tipe 1, generasi, legendary

Referensi

Jure Leskovec, Anand Rajaraman, & Jeffrey D. Ullman (2014)

Mining of Massive Datasets

Cambridge University Press

Alexey Grigorev (16 Agustus 2015)

What is the difference between Manhattan and Euclidean distance measures?

`https://www.quora.com/What-is-the-difference-between-Manhattan-and-Euclidean-distance-measures`

Michael Waskom (2015)

Anscombe's quartet

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

Referensi

David Robinson (6 September 2017)

The Incredible Growth of Python

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

Terima kasih