Normalisasi Data dan Outlier

PERTEMUAN 5
DATA MINING
TIM TEACHING
SISTEM INFORMASI BISNIS
JURUSAN TEKNOLOGI INFORMASI

Outlier

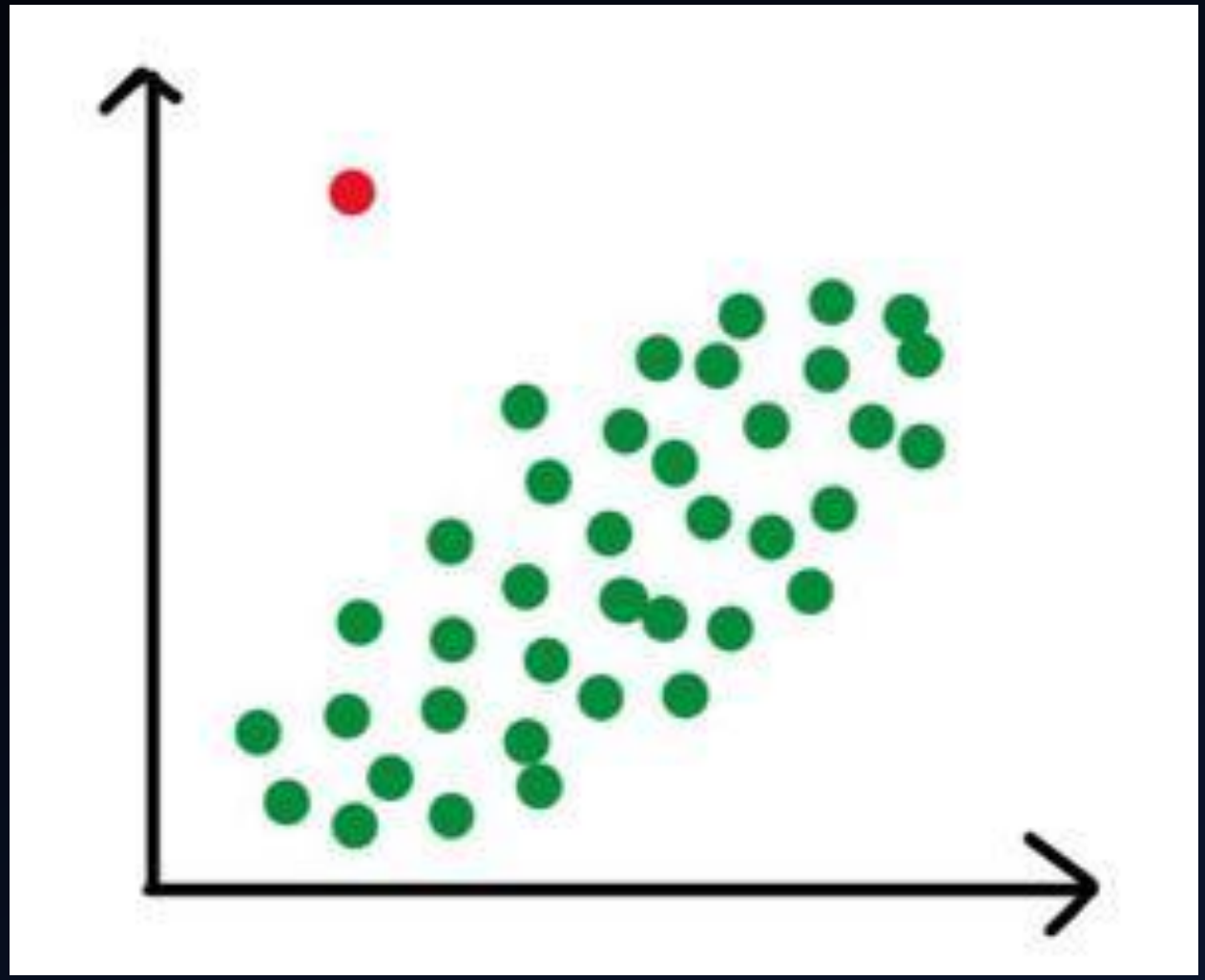
Outlier

- Outlier adalah objek data yang menyimpang secara signifikan dari objek data lainnya dan berperilaku berbeda.
- Outlier adalah suatu objek yang menyimpang secara signifikan dari objek lainnya.
- Hal ini dapat disebabkan oleh kesalahan pengukuran atau pelaksanaan.
- Analisis data outlier disebut dengan outlier analysis atau outlier mining.

Outlier

- Outlier tidak bisa disebut sebagai noise atau error.
- Sebaliknya, mereka diduga tidak dihasilkan dengan metode yang sama seperti objek data lainnya
- Outlier dapat dibagi menjadi 3 bagian, yaitu:
  - Global (or Point) Outliers
  - Collective Outliers
  - Contextual (or Conditional) Outliers

Global Outlier



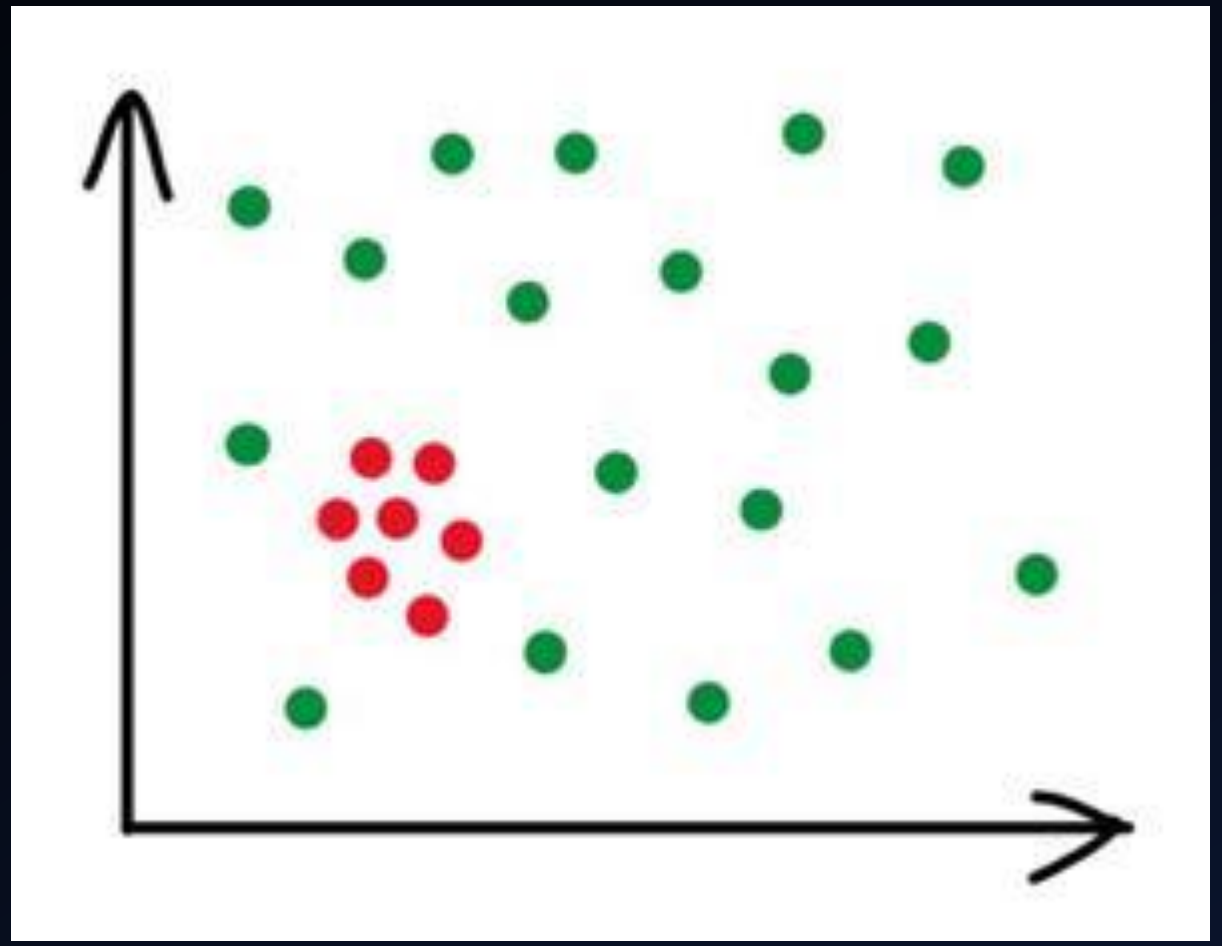
Global Outlier

- Outlier global adalah titik data yang menyimpang secara signifikan dari keseluruhan distribusi kumpulan data
- Kesalahan dalam pengumpulan data, kesalahan pengukuran, atau kejadian yang benar-benar tidak biasa dapat mengakibatkan outlier global
- Outlier global dapat mendistorsi hasil analisis data dan memengaruhi performa model pembelajaran mesin.

Collective Outliers

- Collective Outliers adalah sekelompok titik data yang secara kolektif menyimpang secara signifikan dari keseluruhan distribusi kumpulan data.
- Collective Outliers mungkin bukan outlier jika dipertimbangkan secara individu, namun sebagai sebuah kelompok, mereka menunjukkan perilaku yang tidak biasa.
- Teknik untuk mendeteksi pencilan kolektif mencakup clustering algorithms, density-based methods, dan subspace-based approaches.

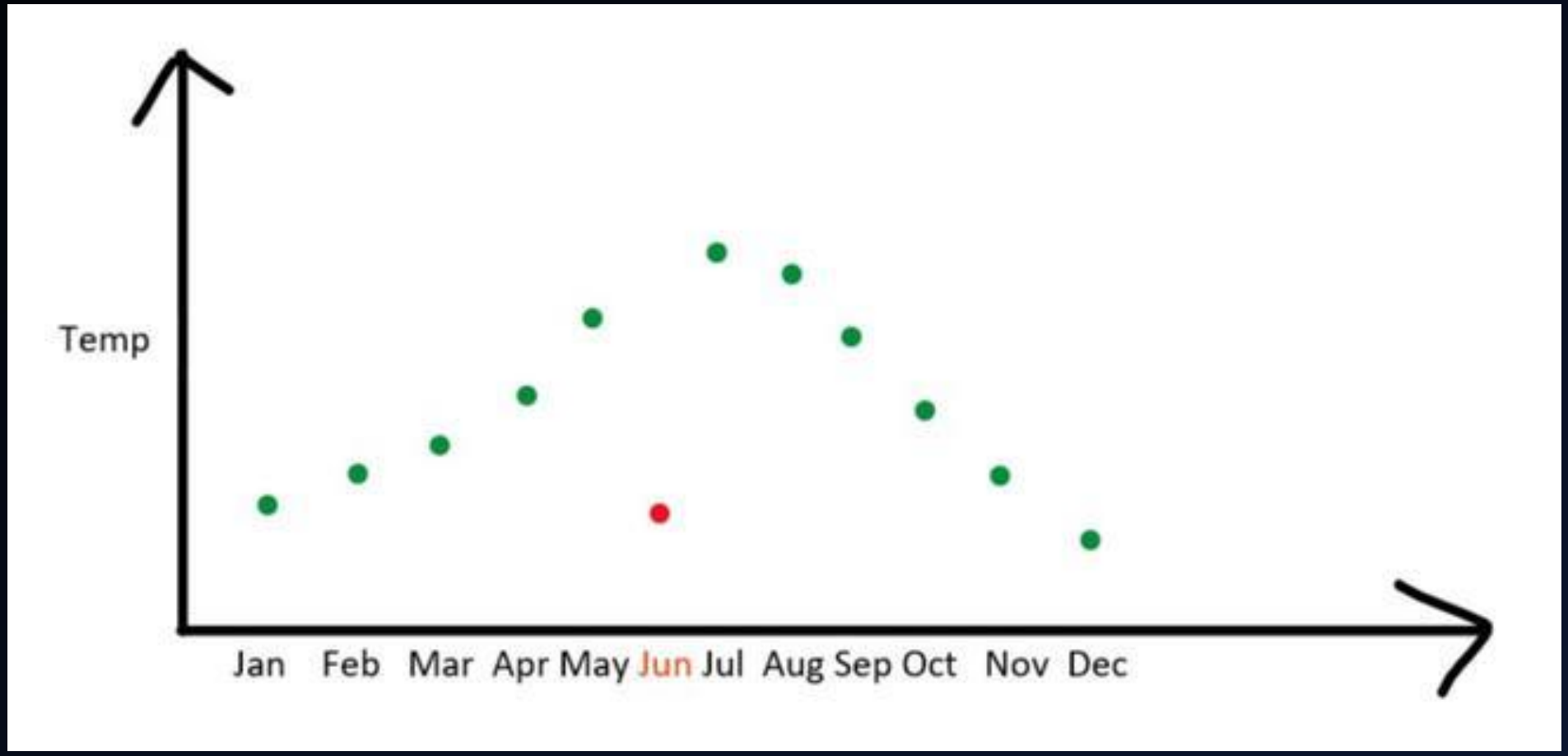
Contoh Collective Outlier



Contextual Outliers

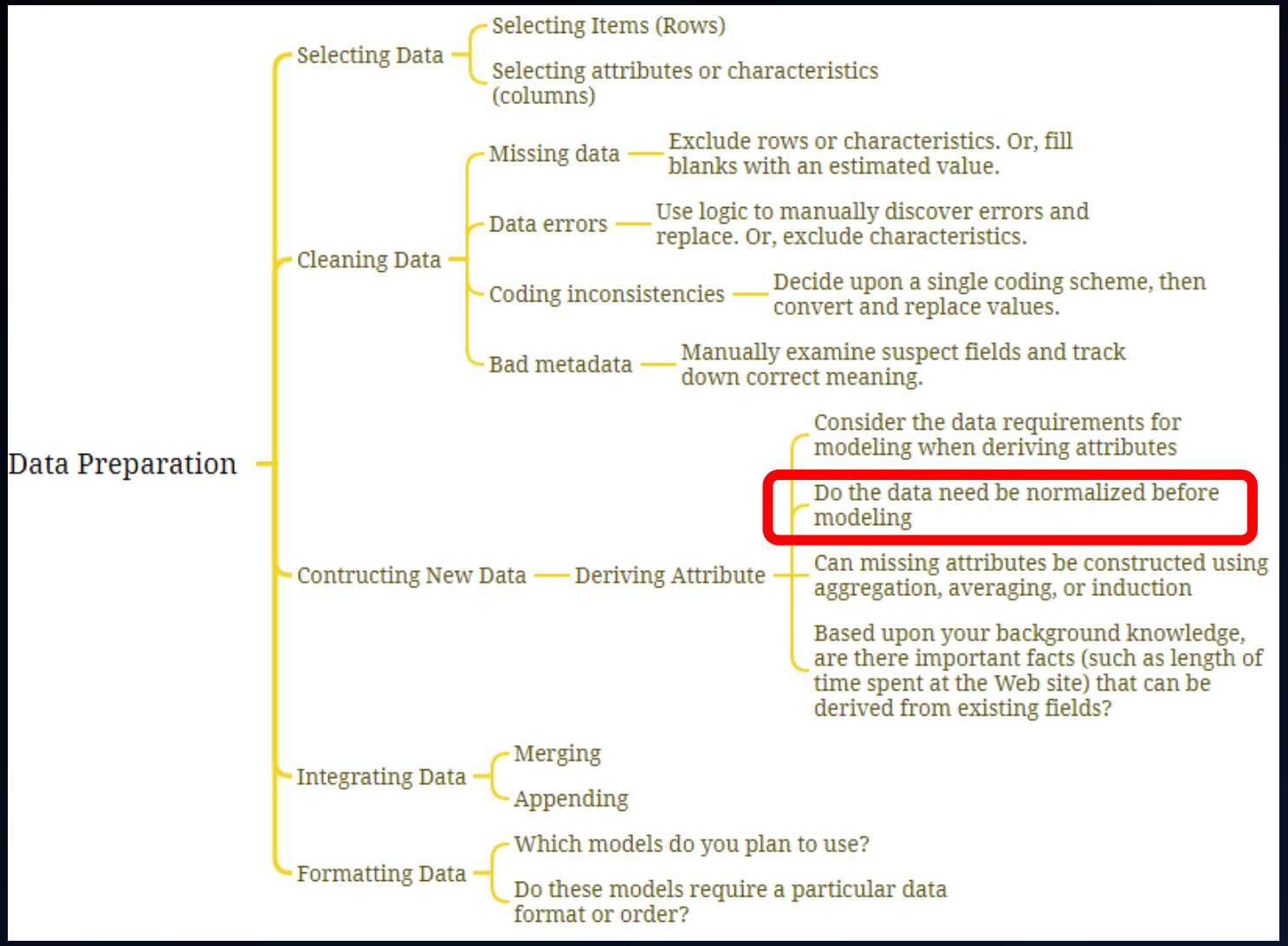
- Contextual Outliers adalah titik data yang menyimpang secara signifikan dari perilaku yang diharapkan dalam konteks atau subkelompok tertentu.
- Contextual Outliers mungkin bukan outlier jika dipertimbangkan dalam keseluruhan kumpulan data, namun menunjukkan perilaku yang tidak biasa dalam konteks atau subgrup tertentu.
- Teknik untuk mendeteksi outlier kontekstual mencakup contextual clustering, contextual anomaly detection, context-aware machine learning.

Contoh Contextual Oulier



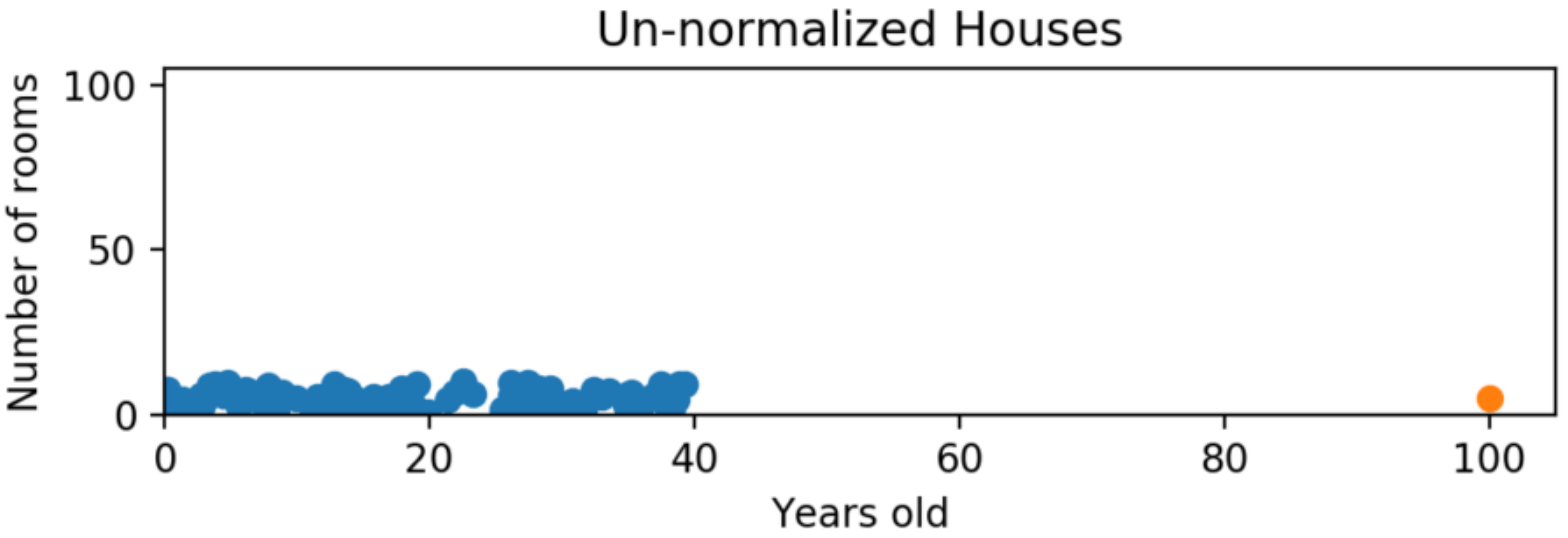
Normalisasi Data

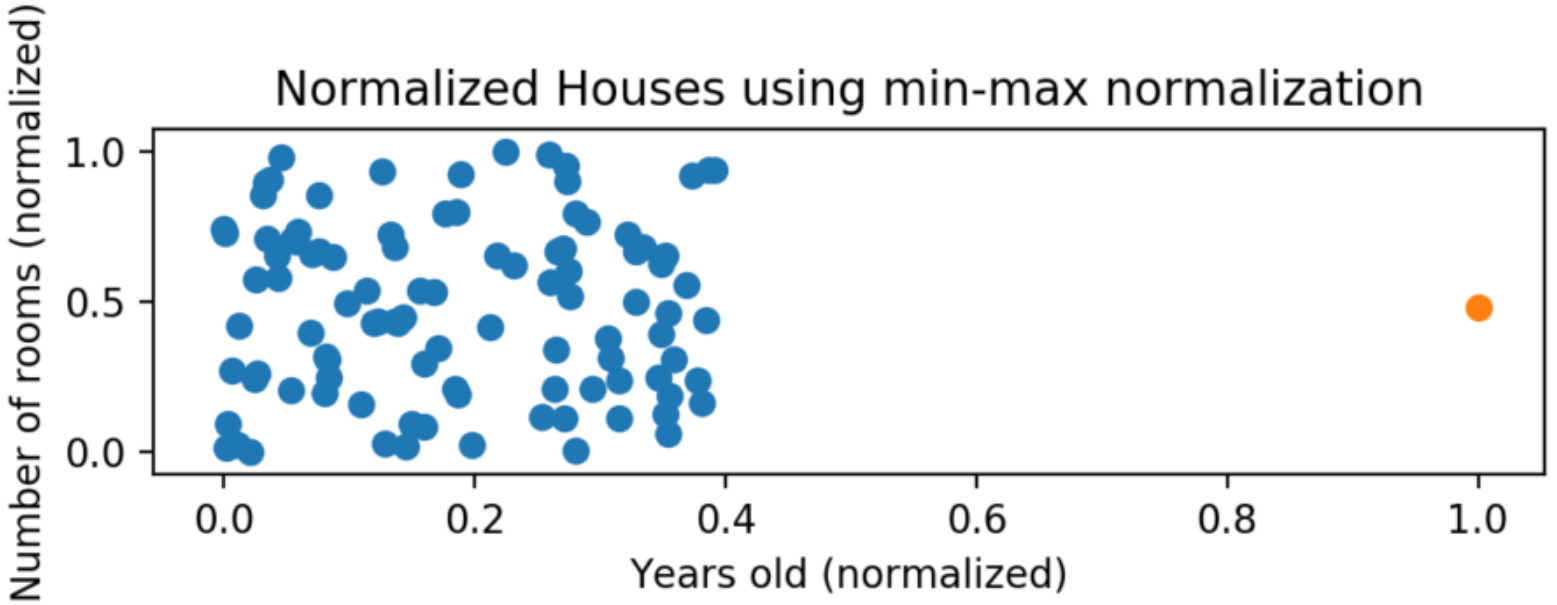
Normalisasi Data



- Normalisasi Data dalam CRISP-DM, termasuk dalam tahapan Data Preparation

Normalisasi Data





Normalisasi Data

- Teknik Normalisasi dalam Data Mining digunakan untuk mengurangi rentang nilai suatu atribut, seperti -1.0 hingga 1.0.
- Normalisasi data terutama digunakan untuk mengurangi data yang berlebihan, sehingga membantu mengurangi ukuran data untuk mempercepat pemrosesan informasi.
- Dalam kebanyakan kasus, Teknik Normalisasi Data dalam Data Mining diimplementasikan dalam model klasifikasi.

Normalisasi Data

- Beberapa manfaat yang didapatkan dengan melakukan Normalisasi Data, antara lain:
  - Menerapkan Data Mining dalam Kumpulan data yang telah ternormalisasi akan lebih mudah
  - Teknik Normalisasi dalam Data Mining yang diterapkan pada kumpulan data yang dinormalisasi memberikan hasil yang lebih akurat dan efektif.
  - Ekstraksi data dari database menjadi lebih cepat setelah data distandarisasi.
  - Pada data yang dinormalisasi, metode analisis data yang lebih khusus dapat digunakan.

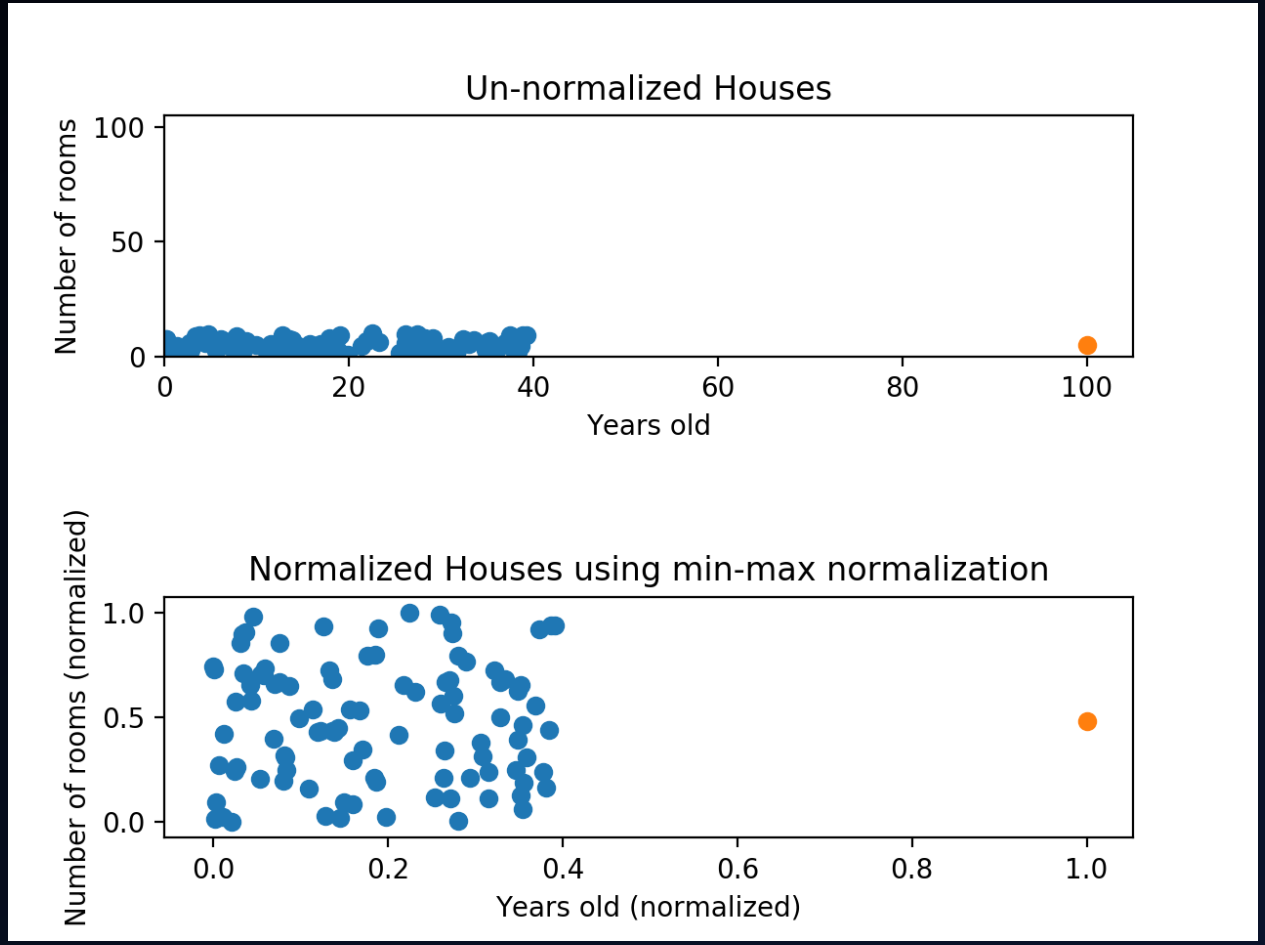
Teknik Normalisasi Data

- Beberapa Teknik normalisasi data yang sering digunakan dalam data mining, antara lain:
  - Min-Max Normalization (Linear Scaling)
  - Z-Score Normalization (Standardization)
  - Decimal Scaling

Min-Max Normalization

- Min-Max Normalization atau yang bisa disebut dengan linear Scaling adalah teknik yang digunakan dalam prapemrosesan data untuk mengubah data numerik menjadi skala umum, biasanya antara 0 dan 1.
- Teknik ini sangat berguna ketika menangani fitur yang memiliki rentang nilai berbeda.
- $x' = \frac{x - \min(X)}{\max(X) - \min(X)}$
- Dimana
  - x' adalah nilai ternormalisasi
  - x adalah nilai asli
  - $\min(X)$ dan $\max(X)$ adalah nilai minimum dan maksimum pada dataset

Visualisasi Min-Max Normalization



Contoh Min-Max Normalization

Student	Math Score	English Score
Alice	85	78
Bob	72	90
Charlie	60	65
David	92	85
Emma	78	80

- Math Score
  - Minimum Math Score ($\min(X_{Math)}$) = 60
  - Maximum Math Score ($\max(X_{Math)}$) = 92
- English Score
  - Minimum English Score ($\min(X_{English)}$) = 65
  - Minimum Math Score ($\max(X_{English)}$) = 90

Contoh Min-Max Normalization

Math Scores:

- Alice: $x' = \frac{85-60}{92-60} = \frac{25}{32} \approx 0.78$
- Bob: $x' = \frac{72-60}{92-60} = \frac{12}{32} = 0.375$
- Charlie: $x' = \frac{60-60}{92-60} = 0$
- David: $x' = \frac{92-60}{92-60} = 1$
- Emma: $x' = \frac{78-60}{92-60} = \frac{18}{32} = 0.5625$

English Scores:

- Alice: $x' = \frac{78-65}{90-65} = \frac{13}{25} = 0.52$
- Bob: $x' = \frac{90-65}{90-65} = 1$
- Charlie: $x' = \frac{65-65}{90-65} = 0$
- David: $x' = \frac{85-65}{90-65} = \frac{20}{25} = 0.8$
- Emma: $x' = \frac{80-65}{90-65} = \frac{15}{25} = 0.6$

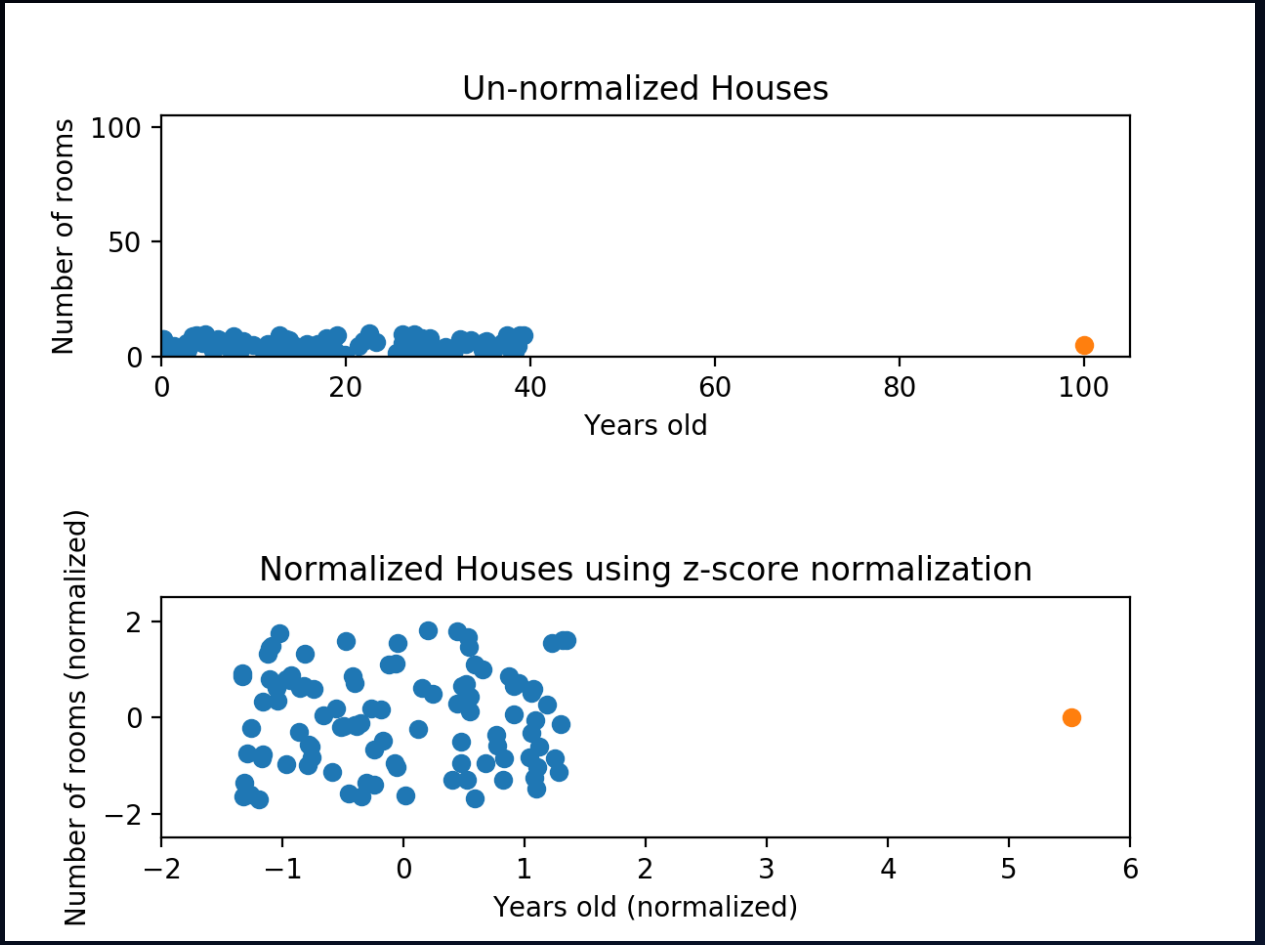
Contoh Min-Max Normalization

- Kesimpulan yang didapatkan
  - Sekarang, skornya telah dinormalisasi antara 0 dan 1.
  - Normalisasi ini memungkinkan kita membandingkan skor antar mata pelajaran yang berbeda tanpa bias pada skala yang berbeda.
  - Misalnya, skor 0,78 dalam Matematika untuk Alice menunjukkan bahwa dia mendapat skor 78% dari selisih antara skor Matematika minimum dan maksimum dalam kumpulan data.
  - Demikian pula, skor 0,52 dalam bahasa Inggris untuk Alice menunjukkan bahwa dia mendapat skor 52% dari selisih antara skor bahasa Inggris minimum dan maksimum.

Z-Score Normalization

- Z-Score Normalization, disebut juga Standard Score Normalization atau Standardization teknik statistik yang digunakan untuk mengubah data menjadi distribusi normal standar dengan mean 0 dan standar deviasi 1.
- $z = \frac{x - \mu}{\sigma}$
- Dimana
  - z adalah nilai ternormalisasi
  - x adalah nilai data asli
  - μ adalah nilai Means (rata-rata)
  - σ adalah standar deviasi, didapatkan dari $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Visualisasi Z-Score Normalization



Contoh Z-Score Normalization

Employee	Performance Rating
Alice	85
Bob	72
Charlie	60
David	92
Emma	78

- Mean (μ) = $\frac{85+72+60+92+78}{5} = \frac{387}{5} = 77.4$
- Standard Deviation (σ) ≈ 11.67

For Alice:

$$z_{\text{Alice}} = \frac{85-77.4}{11.67} \approx \frac{7.6}{11.67} \approx 0.65$$

For Bob:

$$z_{\text{Bob}} = \frac{72-77.4}{11.67} \approx \frac{-5.4}{11.67} \approx -0.46$$

For Charlie:

$$z_{\text{Charlie}} = \frac{60-77.4}{11.67} \approx \frac{-17.4}{11.67} \approx -1.49$$

For David:

$$z_{\text{David}} = \frac{92-77.4}{11.67} \approx \frac{14.6}{11.67} \approx 1.25$$

For Emma:

$$z_{\text{Emma}} = \frac{78-77.4}{11.67} \approx \frac{0.6}{11.67} \approx 0.05$$

Contoh Z-Score Normalization

- Kesimpulan
  - Skor z positif menunjukkan peringkat di atas rata-rata, sedangkan skor z negatif menunjukkan peringkat di bawah rata-rata. Skor z 0 berarti peringkatnya tepat pada mean.
    - Peringkat kinerja Alice berada di atas rata-rata yang ditunjukkan dengan z-score positif sebesar 0,65.
    - Peringkat kinerja Bob berada di bawah rata-rata yang ditunjukkan dengan z-score negatif sebesar -0,46.
    - Peringkat kinerja Charlie berada jauh di bawah rata-rata yang ditunjukkan dengan z-score negatif sebesar -1,49.
    - Peringkat kinerja David jauh di atas rata-rata yang ditunjukkan dengan z-score positif sebesar 1,25.
    - Penilaian kinerja Emma sangat mendekati mean yang ditunjukkan dengan z-score sebesar 0,05.

Decimal Scaling

- Decimal Scaling adalah teknik normalisasi data yang melibatkan pergeseran titik desimal dari setiap titik data ke kiri atau kanan untuk membawa nilai dalam rentang yang diinginkan.
- Metode ini sangat berguna ketika data asli mencakup rentang besaran yang luas.
- Dengan menskalakan data dengan cara ini, dipertahankan perbedaan relatif antara titik-titik data sambil memastikan titik-titik tersebut berada dalam rentang yang ditentukan.

Decimal Scaling

- Cara kerja Decimal Scaling:
  - **Determine the Scaling Factor,** Pilih faktor skala, yang biasanya pangkat 10. Faktor skala menentukan seberapa banyak Anda menggeser koma desimal
  - **Normalize the Data,** untuk setiap titik data, bagilah dengan faktor skala yang dipilih. Tindakan ini menggeser koma desimal ke kiri jika faktor skala lebih besar dari 1 atau ke kanan jika faktor skala kurang dari 1.
- $x' = \frac{x}{10^j}$
- Dimana
  - x' adalah nilai ternormalisasi
  - x adalah nilai data asli
  - j adalah Jumlah tempat desimal yang diperlukan untuk mewakili nilai terbesar dalam kumpulan data.

Contoh Decimal Scaling

Month	Product A	Product B	Product C
January	5000	75000	100000
February	6000	80000	110000
March	5500	70000	105000

- Nilai maksimum adalah 110000
- Nilai Faktor Skala adalah 100000

For Product A:

- January: $x' = \frac{5000}{100000} = 0.05$
- February: $x' = \frac{6000}{100000} = 0.06$
- March: $x' = \frac{5500}{100000} = 0.055$

For Product B:

- January: $x' = \frac{75000}{100000} = 0.75$
- February: $x' = \frac{80000}{100000} = 0.8$
- March: $x' = \frac{70000}{100000} = 0.7$

For Product C:

- January: $x' = \frac{100000}{100000} = 1$
- February: $x' = \frac{110000}{100000} = 1.1$
- March: $x' = \frac{105000}{100000} = 1.05$