

A decorative background featuring a network diagram. It consists of numerous nodes, represented by small circles, connected by thin lines. Some nodes are solid blue, while others are outlined in blue. The network is more densely packed on the left and right sides of the slide, with the central area being mostly white space containing the title.

Handling Missing Data

Penyebab *missing value*?

1. Data Extraction
2. Data Collection:
 - a. Lost / Data yang hilang: misal data yang tidak tersimpan dengan sempurna
 - b. Not Exist: misal data yang tercipta dari pembagian dua variabel dengan nilai pembagi = 0
 - c. Not Found: misal data yang direferensikan ternyata tidak pernah ada (salah alamat, salah nomor telpon, dsb)



Dampaknya?

- Tidak compatible dengan *Scikit-learn*
- Terdapat kemungkinan adanya imputasi data mengganggu distribusi variabel
- Berdampak pada Model Machine Learning

Variasi

- ⊙ Missing completely at random (MCAR)
- ⊙ Missing at random (MAR)
- ⊙ Missing not at random (MNAR)

MCAR



- ⦿ Probabilitas data hilang adalah sama untuk semua pengamatan
- ⦿ Tidak ada hubungan antara data yang hilang dan nilai lainnya, yang diamati atau hilang, dalam kumpulan data

MAR

- ◎ Data hilang pada tingkat tertentu tetapi tingkat itu tergantung pada beberapa variabel lain dalam data.
- ◎ MAR terjadi ketika ada hubungan antara kecenderungan nilai yang hilang dan data yang diamati. Dengan kata lain, probabilitas observasi yang hilang tergantung pada informasi yang tersedia (yaitu, variabel lain dalam dataset).

MNAR

- ⊙ Ada hubungan antara kecenderungan nilai yang akan hilang dan nilai-nilai dari variabel tersebut. Dengan kata lain, data MNAR terjadi ketika nilai yang hilang dari suatu variabel terkait dengan nilai variabel itu sendiri, bahkan setelah mengendalikan variabel lain.
- ⊙ Terjadi ketika ada mekanisme atau alasan mengapa nilai yang hilang dimasukkan ke dalam kumpulan data



Missing Data Imputation

Missing Data Imputation

- ◎ Imputasi adalah tindakan mengganti data yang hilang dengan perkiraan statistik dari nilai yang hilang.
- ◎ Tujuan dari setiap teknik imputasi adalah untuk menghasilkan kumpulan data lengkap yang dapat digunakan untuk melatih model ML.

Imputation Method

1. Complete Case Analysis (CCA)
2. Mean or Median Imputation
3. Arbitrary Value Imputation
4. End of Tail Imputation

Complete Case Analysis

Merupakan cara imputasi dengan membuang pengamatan di mana nilai dalam salah satu variabel hilang.

Dengan CCA, pengamatan yang akan dianalisis merupakan pengamatan yang ada informasinya di semua variabel dalam kumpulan data.

Gender	Price	Make	Engine
Female	100	Ford	2000
	90	Ford	2000
Male	50	Kia	1500
Male	60	Kia	
Female	120	Nissan	3000
Female		BMW	4500
Male	200	BMW	4500



Gender	Price	Make	Engine
Female	100	Ford	2000
Male	50	Kia	1500
Female	120	Nissan	3000
Male	200	BMW	4500

- Observations with missing values are removed

Complete Case Analysis

Kelebihan:

- Mudah diterapkan
- Tidak diperlukan manipulasi data
- Mempertahankan distribusi variabel (jika data adalah MCAR, maka distribusi variabel dari kumpulan data yang direduksi harus sesuai dengan distribusi dalam kumpulan data asli)

Kekurangan:

- Dapat berpotensi mengecualikan sebagian besar dari dataset asli (jika data yang hilang berlimpah)
- Pengamatan yang dikecualikan bisa menjadi informasi untuk analisis (jika data tidak hilang secara acak)
- CCA akan membuat dataset bias jika kasus lengkap berbeda dari data asli (misalnya, ketika informasi yang hilang adalah MAR atau NMAR dan tidak hilang secara acak).

Complete Case Analysis

Kapan menggunakan CCA?

1. Data MCAR
2. Missing value maksimal ~5% dari dataset

Mean or Median Imputation

Imputasi
mean/median terdiri
dari mengganti
semua kemunculan
nilai yang hilang (NA)
dalam suatu variabel
dengan mean atau
median.

Price		Price
100	Mean = 86.66 Median = 90 ➔	100
90		90
50		50
40		40
20		20
100		100
		86.66
60		60
120		120
		86.66
200		200

Mean or Median Imputation

Kelebihan:

- Mudah diimplementasikan.
- Cara cepat untuk mendapatkan dataset lengkap.

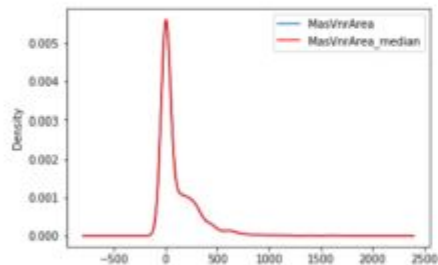
Kekurangan:

- Distorsi distribusi variabel asli.
- Distorsi varians asli.

Mean or Median Imputation

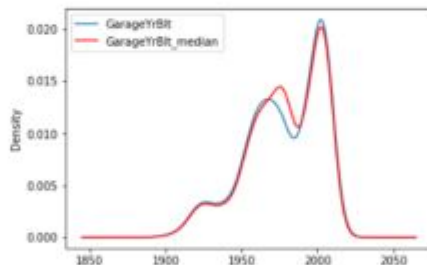
Mean / Median Imputation effects

MasVnrArea 0.5% missing obs



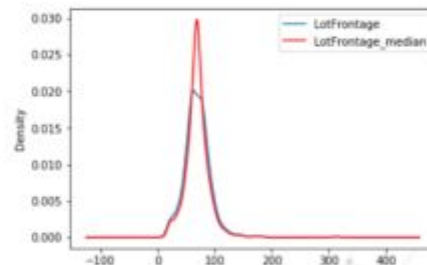
Variance: 32983
Variance after imputation: 32874

GarageYrBlt 5.5% missing obs



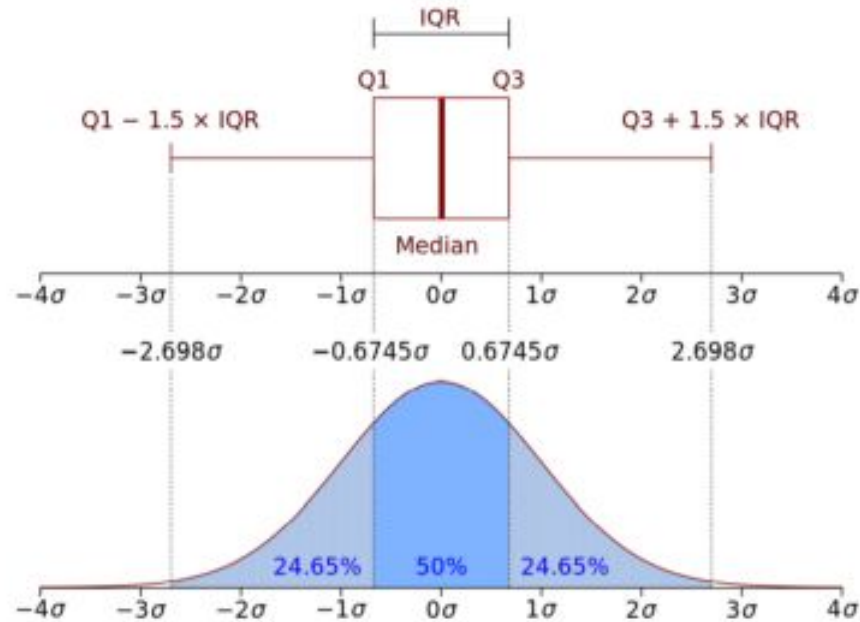
Variance: 624
Variance after imputation: 591

LotFrontage 17% missing obs



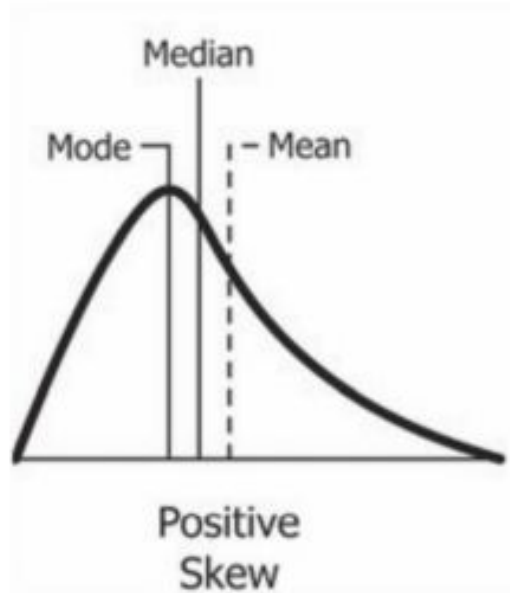
Variance: 532
Variance after imputation: 434

Mean or Median Imputation



- If the variable is normally distributed the mean and median are approximately the same

Mean or Median Imputation



- If the variable is skewed, the median is a better representation

Mean or Median Imputation

Kapan menggunakan Mean or Median Imputation?

1. Data MCAR*
2. Missing value maksimal ~5% dari dataset

*Meskipun secara teori, kondisi di atas harus dipenuhi untuk meminimalkan dampak dari teknik imputasi ini, dalam praktiknya, imputasi mean/median sangat umum digunakan, bahkan dalam kasus-kasus ketika data tidak MCAR dan ada banyak nilai yang hilang. Alasan di balik ini adalah kesederhanaan teknik.

Arbitrary Value Imputation

- Metode imputasi dengan mengganti semua kemunculan nilai yang hilang (NA) dalam variabel dengan nilai arbitrer.
- Biasanya nilai arbitrer yang digunakan adalah 0, 999, -999 (atau kombinasi lain dari 9) atau -1 (jika distribusinya positif).
- Cocok untuk variabel numerik dan kategoris

Arbitrary Value Imputation

Price
100
90
50
40
20
100
60
120
200

~~Arbitrary~~ = 99



Price
100
90
50
40
20
100
999
60
120
999
200

Asumsi:

- Data MNAR

Arbitrary Value Imputation

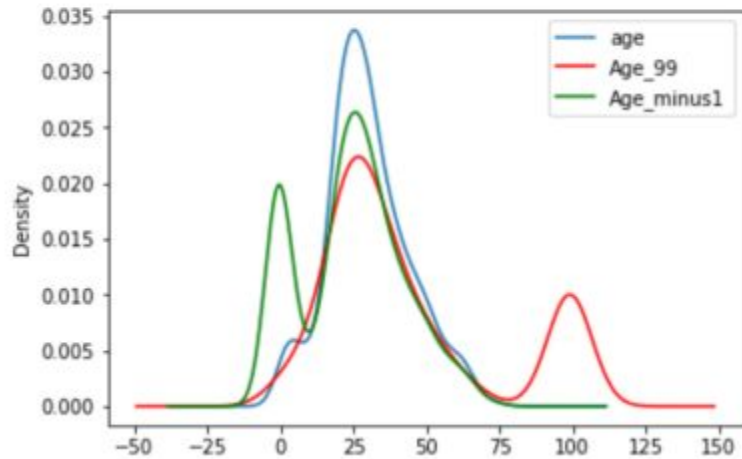
Kelebihan:

- Mudah diterapkan
- Cara cepat untuk mendapatkan kumpulan data lengkap

Kekurangan:

- Distorsi distribusi variabel asli.

Arbitrary Value Imputation



- ~20% of data is missing in Age

Original variable variance: 194
Variance after 99 imputation: 888
Variance after -1 imputation: 307

Arbitrary Value Imputation

Kapan menggunakan Arbitrary Value Imputation?

Mengganti NA dengan nilai arbitrer dapat digunakan ketika ada alasan untuk percaya bahwa NA tidak hilang secara acak. Mengganti data NA dengan median atau mean akan membuat NA terlihat seperti sebagian besar pengamatan lainnya.

End of Tail Imputation

- Metode ini setara dengan metode Arbitrary Value Imputation, tetapi secara otomatis memilih nilai arbitrer di akhir distribusi variabel.
- Jika variabel terdistribusi normal, maka menggunakan mean plus atau minus 3 kali standar deviasi.
- Jika variabel nya berbentuk *skewed*, maka menggunakan aturan kedekatan IQR.
- Lebih sesuai untuk variabel numerik.

End of Tail Imputation

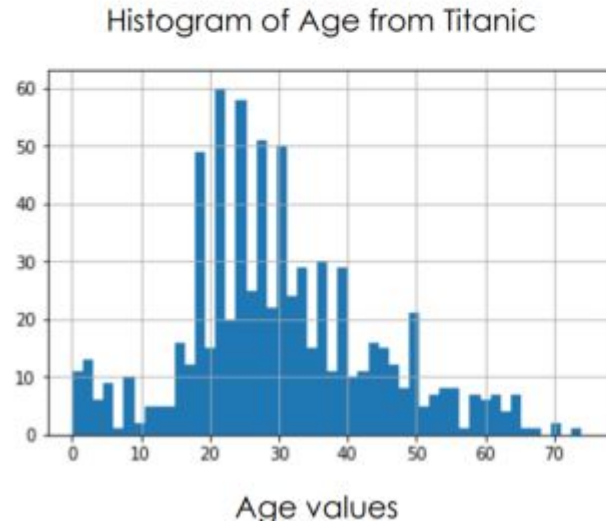
Kelebihan:

- ⊙ Mudah diterapkan
- ⊙ Cara cepat untuk mendapatkan kumpulan data lengkap

Kekurangan:

- ⊙ Distorsi distribusi variabel asli
- ⊙ Distorsi varians asli
- ⊙ Distorsi kovarians dengan variabel yang tersisa dari kumpulan data
- ⊙ Teknik ini dapat menutupi outlier yang sebenarnya dalam distribusi

End of Tail Imputation

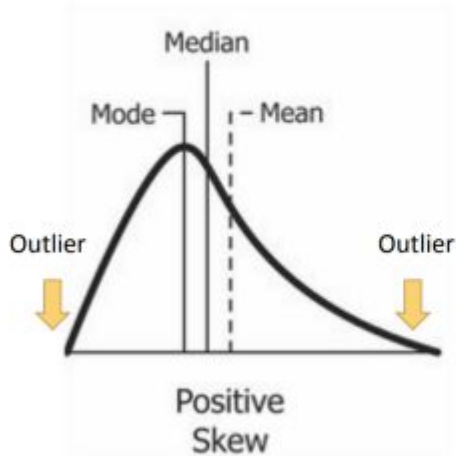


Mean(Age) + 3 × std(Age) = 72



End of Tail Imputation

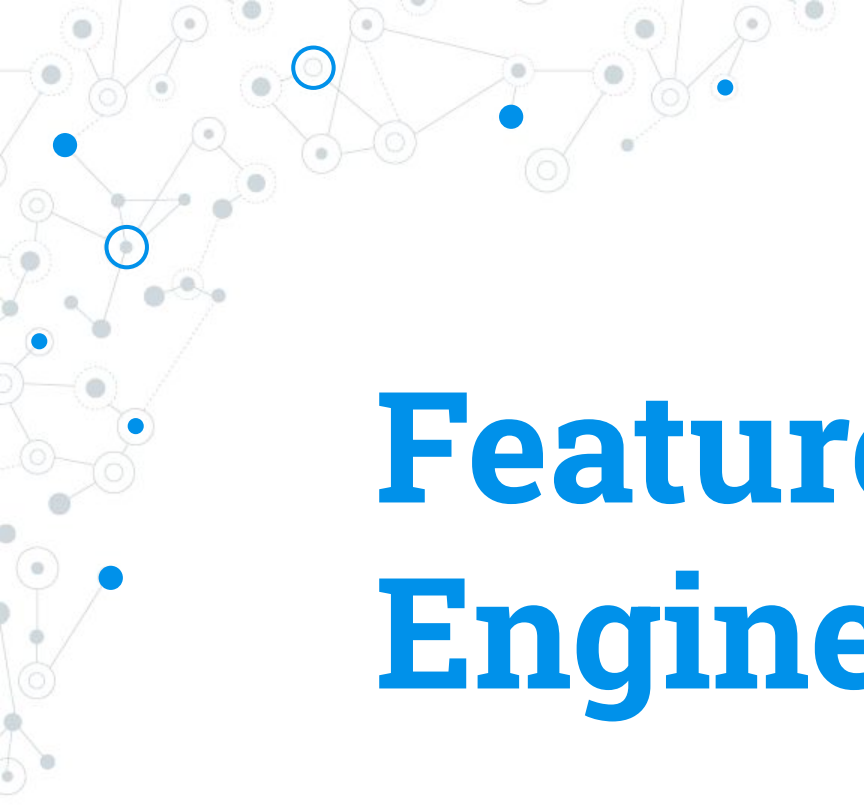
- **Skewed distributions**




- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:

- $IQR = 75^{\text{th}} \text{ Quantile} - 25^{\text{th}} \text{ Quantile}$
- $\text{Upper limit} = 75^{\text{th}} \text{ Quantile} + IQR \times 1.5$
- $\text{Lower limit} = 25^{\text{th}} \text{ Quantile} - IQR \times 1.5$

Note, for extreme outliers, multiply the IQR by 3 instead of 1.5

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with blue dots. The lines are thin and gray, creating a mesh-like structure.

Feature Engineering

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a cluster of nodes connected by lines, with several nodes highlighted in blue.

Feature Engineering?

Proses untuk mengubah data mentah menjadi *feature* (karakteristik, atribut, dll) agar bisa merepresentasikan data yang lebih baik dalam model, sehingga bisa menambah tingkat akurasi dari model yang dibuat

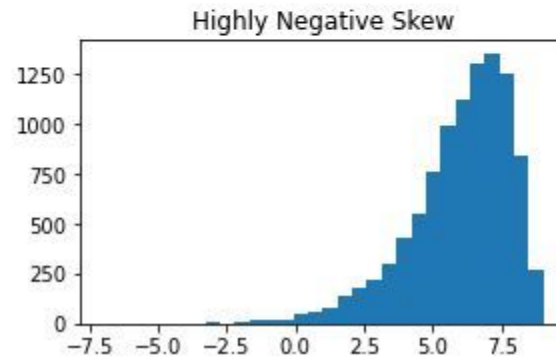
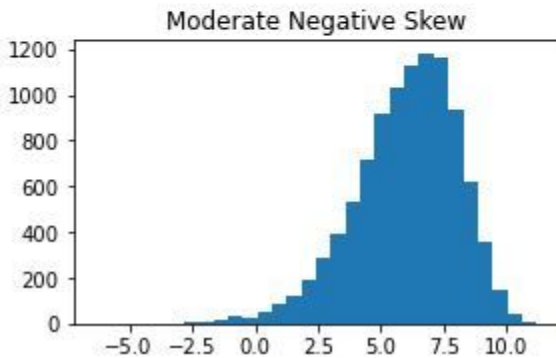
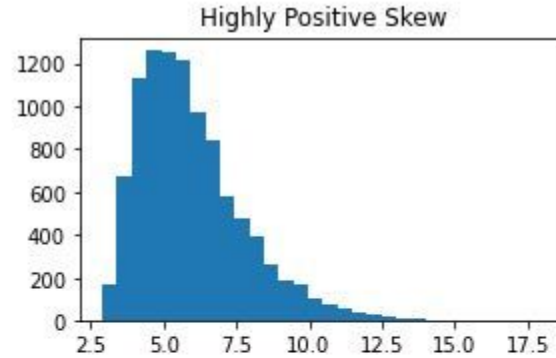
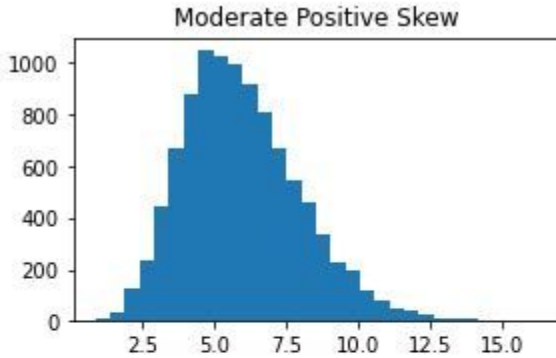


A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

1.

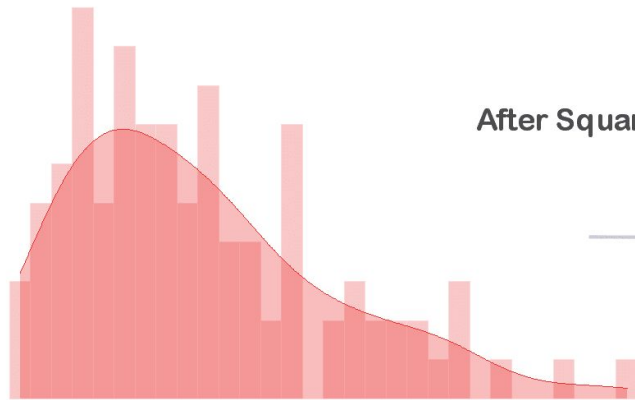
Data Transformation

Data Transformation

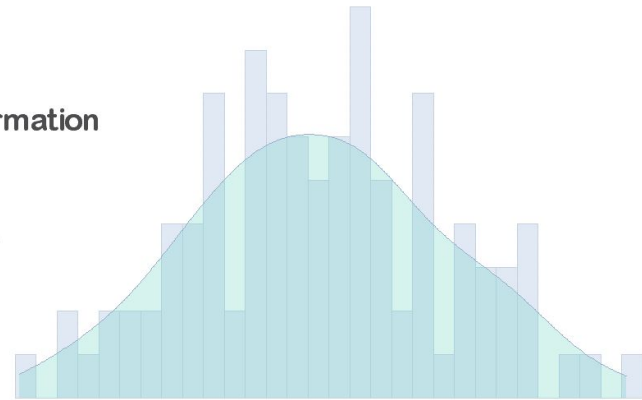


Beberapa teknik transformasi yang bisa digunakan:

- ◎ Root transformation
- ◎ Log transformation
- ◎ Square/Power transformation



After Square Root Transformation





2.

Data Normalization

Tanpa Normalisasi

Umur	Gaji
20	100000
30	20000
40	500000
...	...



Dengan Normalisasi

Umur	Gaji
0.2	0.2
0.3	0.04
0.4	1
...	...

Beberapa teknik yang sering digunakan:

Min-max normalization

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Metode ini menjadikan nilai terkecil dari sebuah fitur menjadi 0, dan nilai terbesar fitur menjadi 1, kemudian membagi jarak antar nilai secara proporsional

Z-score normalization

$$x_{new} = \frac{x_{old} - \mu}{\sigma}$$

Metode ini menjadikan nilai rata-rata sebuah fitur menjadi 0 dan standar deviasi menjadi 1





3. **Binning**

Apa itu binning?

Binning adalah sebuah proses untuk mengelompokkan data ke dalam bagian-bagian yang lebih kecil yang disebut **bin** berdasarkan kriteria tertentu

Age 0-15 = child

Age 16-29=young adult

Age 20-50 = adult

Age >50 = elderly

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting different levels or types of nodes. The lines are thin and gray, connecting the nodes in a non-linear fashion.

4. Encoding

Kenapa?

Karena komputer tidak dapat memproses data bertipe kategori sehingga kita harus mengubah data tersebut menjadi berbentuk bilangan.

One-Hot Encoding

One-Hot encoding adalah salah satu metode encoding. Metode ini merepresentasikan data bertipe kategori sebagai vektor biner yang bernilai integer, 0 dan 1, dimana semua elemen akan bernilai 0 kecuali satu elemen yang bernilai 1, yaitu elemen yang memiliki nilai kategori tersebut.



ID	Jakarta	Bandung	Surabaya	Pontianak	Medan	Makassar	Jayapura
100	1	0	0	0	0	0	0
101	0	1	0	0	0	0	0
102	0	0	1	0	0	0	0
103	0	0	0	1	0	0	0
104	0	0	0	0	1	0	0
105	0	0	0	0	0	1	0
106	0	0	0	0	0	0	1



Thanks!

Any questions?