KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 14:

Outlier

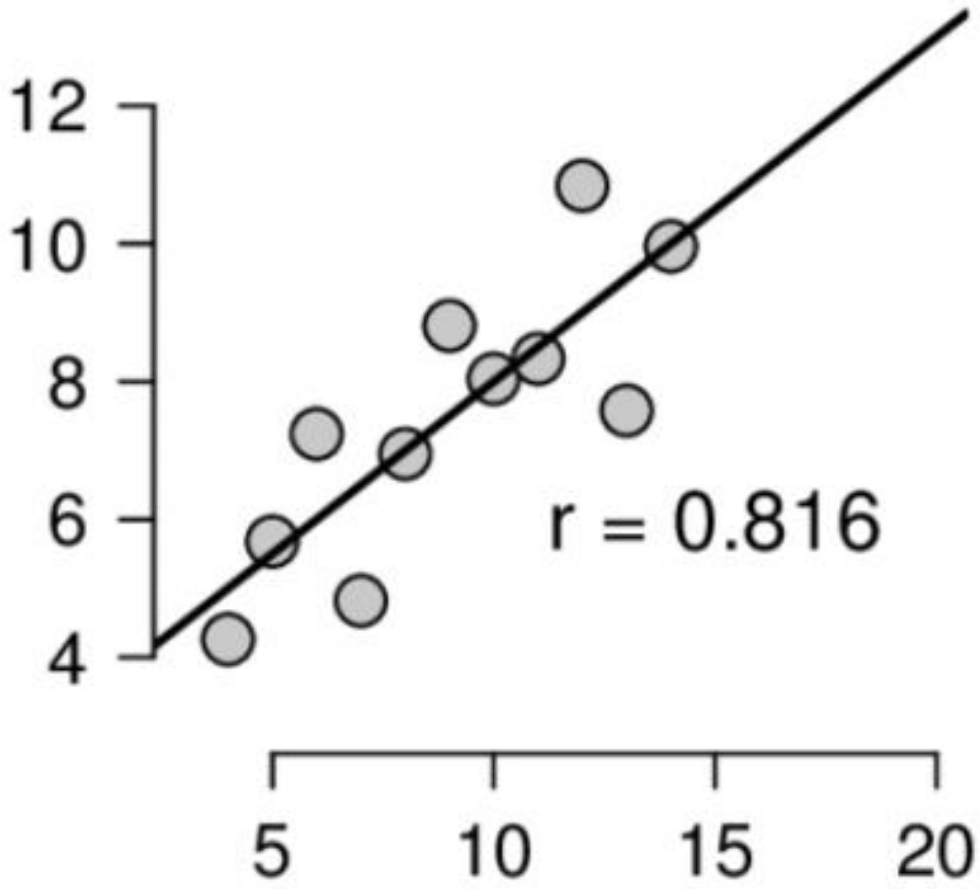
Outline

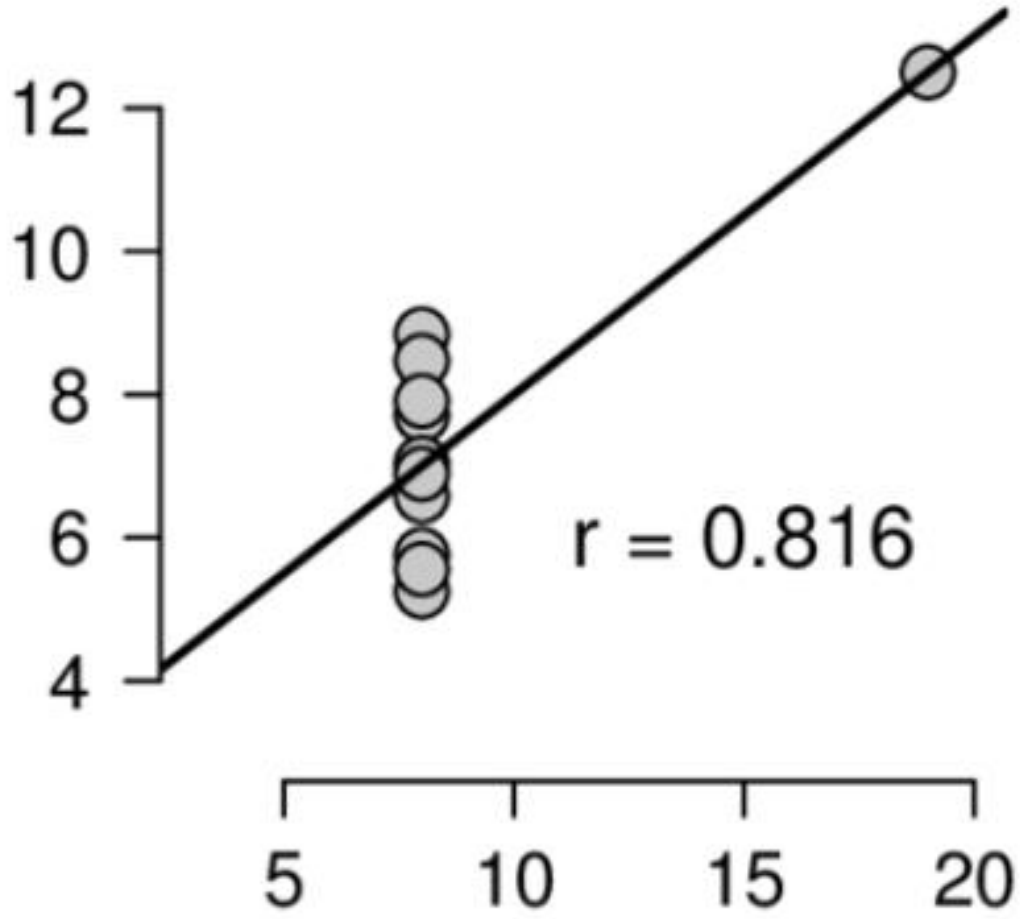
- Definisi data outlier
- Kriteria dan karakteristik data outlier
- Cara mengidentifikasi data outlier
- Cara mengatasi data outlier

Apa itu Outlier?

- Sebuah data yang memiliki karakteristik berbeda dengan data lain dalam satu dataset yang sama
- Data yang dianggap menyimpang
- Data outlier biasanya memiliki nilai ekstrim, dan berada di luar distribusi data
- Missing values hampir mirip dengan data outlier
- Data outlier dapat mengaburkan hasil analisis data
- Namun, terkadang data outlier adalah insight yang dicari.

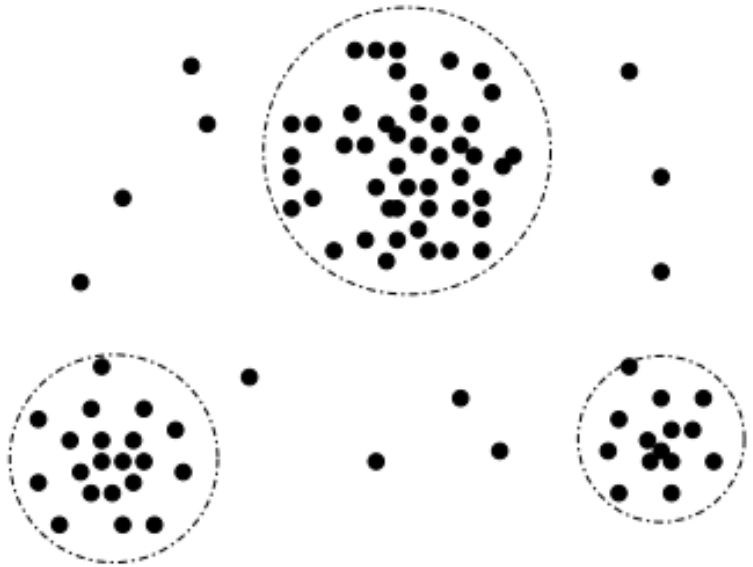
Efek data Outlier pada Regresi Linier





Data Outlier yang Dicari

- Beberapa kasus seperti fraud detection, malah mencari data outlier
- Hal ini karena banyaknya transaksi normal dan sedikitnya fraud
- Data transaksi normal yang dicluster, akan berkelompok
- Data outlier akan tersisih dan terdeteksi sebagai fraud



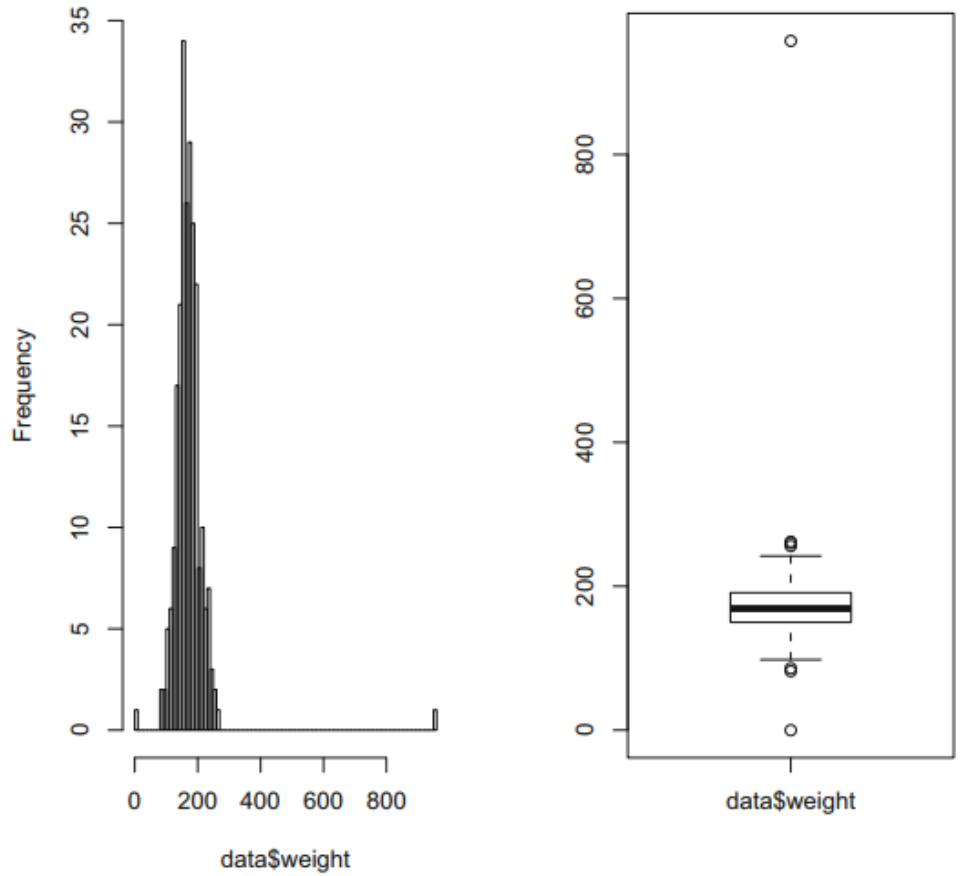
Mendeteksi Outlier

- Outlier sulit dideteksi menggunakan summary statistics
- Hal ini dikarenakan outlier dianggap data yang setara, dan pengukuran central tendency tidak dapat mendeteksi data outlier
- Melalui histogram dan/atau boxplot, outlier dapat terlihat
- Contoh: data survey pada package MASS

```
> summary(data)
     sex          height          weight      handedness  exercise
 Female:118   Min.   :59.00   Min.   :  0.0   Left : 18   Freq:115
 Male  :118   1st Qu.:65.27   1st Qu.:150.0   Right:218   None: 24
 NA's  :  1   Median :67.00   Median :169.0   NA's :  1   Some: 98
              Mean   :67.85   Mean   :172.9
              3rd Qu.:70.42   3rd Qu.:191.0
              Max.   :79.00   Max.   :958.0
   smoke
 Heavy: 11
 Never:190
 Occas: 19
 Regul: 17
```

cont’d

Dari summary statistics, tidak mudah melihat outlier



Histogram & Boxplot

- Dapat dengan mudah melihat outlier
- Pada histogram, terlihat data yang berada diluar distribusi
- Pada boxplot, dilambangkan dengan bulatan diluar batas whisker (1.5 x InterquartileRange)

Data Cleaning

- Sebuah proses umum yang dilakukan sebelum analisis data lebih lanjut
- Untuk mengatasi problem missing values dan data outlier (yang merugikan)
- Sering terintegrasi pada preprocessing data di proses ETL

Data Cleaning Methods

- Beberapa metode untuk mengatasi outlier dan missing values:
- Hapus row dimana terdapat missing value atau outlier pada atribut
- Isi missing value secara manual
- Gunakan konstanta universal
- Gunakan pengukuran tendensi tengah (mean atau median) untuk mengisi missing value
- Gunakan nilai yang paling memungkinkan
- Penggunaan metode tersebut memiliki kerugian dimana data menjadi bias dan bisa menghilangkan informasi yang dicari.

Tugas

- Melakukan identifikasi dan mengatasi adanya data outlier pada dataset berikut:
  - dataset mpg (package ggplot2) khusus variabel highway miles per gallon
  - dataset warpbreaks
- Mencatat karakteristik serta penyebab adanya data outlier.
- 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 14_Kelompok XX. (Contoh: Tugas 14_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