Data Science with R

Detecting outlier and Missing Value- Preprocessing Data

Statistics – ITS

Regita Putri Permata

(Master Student at Statistics Department – ITS)

Meet 2: Preprocessing data with R

Outline

- EDA (Statistics Descriptive)
- Data Cleaning(missing data, noisy data)
- Data Transformation

Why Preprocess the Data ?

- **real-world databases are** highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources
- Incomplete data

  lacking attribute values or certain attributes of interest, or containing only aggregate data. e.g., sales=“ ”
- Innacurate or Noisy

  containing errors, or values that deviate from the expected.
  e.g., age=“-10
- Inconsistent

  containing discrepancies in the codes or names. e.g., Was rating “1,2,3”, now rating “A, B, C”

Why is Data Processing Important?

- No quality data, no quality mining results
    - Quality decisions must be based on quality data
      e.g., duplicate or missing data may cause incorrect or even misleading statistics.
    - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse. —Bill Inmon

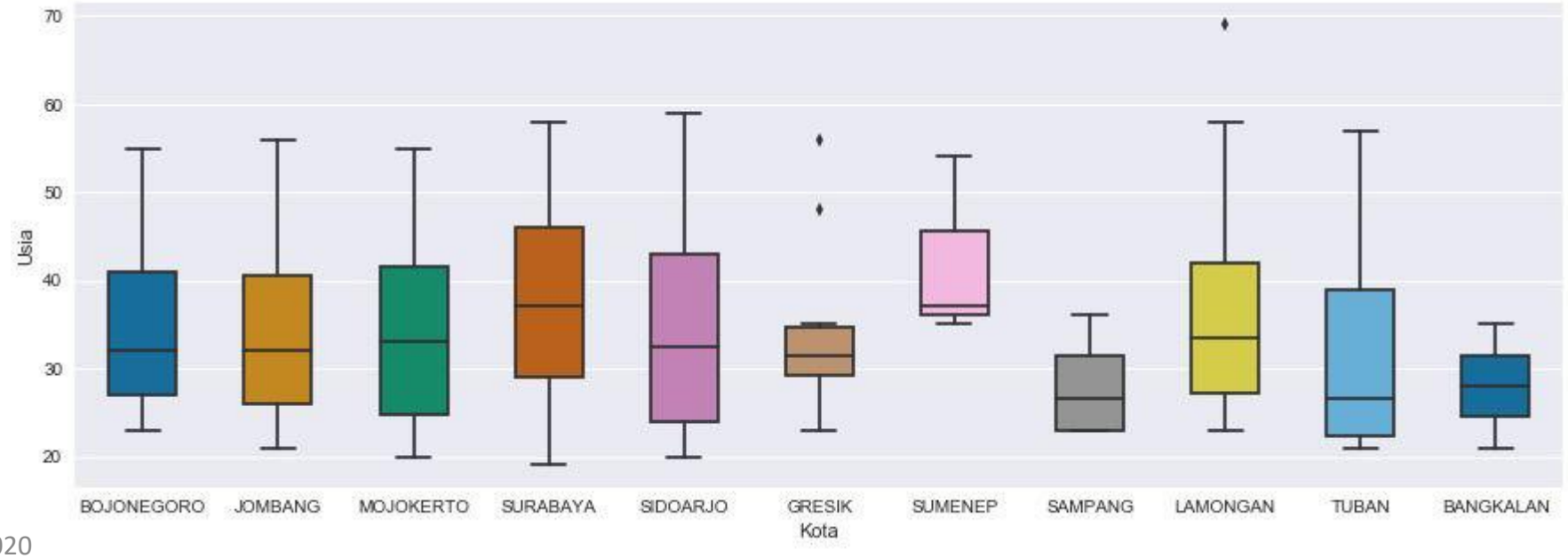
Explanatory Data Analysis

Descriptive Statistics

- **Purpose:**
    - Knowing whether there is a missing value or not
    - Find out whether there are outliers or not
    - Find out if there are different formats
- **Visualization:**
    - know the characteristics / data patterns
    - Knowing whether there are outliers or not using plots
    - Knowing the distribution of data

Outlier

- Outliers are different data / outliers and commonly called anomalies where outliers have meaning in a data. **Outliers do not equal errors**. They should be detected, but not necessarily removed. Their inclusion in the analysis is a statistical decision.
- Noise is a different data where the data has no meaning. For example format errors in inputting.



Data Cleaning

Data Cleaning

- Data Cleaning is the process of transforming raw data **into consistent data** that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability.
- **Data cleaning** routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- Typical actions like imputation or outlier handling obviously influence the results of a statistical analyses
- Consistent data is the stage where data is ready for statistical inference. It is the data that most statistical theories use as a starting point

Missing data

- Data is not available
    - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- A missing data/value, represented by NA in R, is a placeholder for a datum of which the type is known but its value isn’t.
- Therefore, it is impossible to perform statistical analysis on data where one or more values in the data are missing.

How to Handle Missing data ?

- Overcoming missing value depends on the data. If the data is categorical, the input is better to use **mode**, if the data is continuous, it is better to use the **mean or median**

```
>age <- c(23, 16, NA)
> mean(age)
[1] NA
> mean(age, na.rm = TRUE)
[1] 19.5
```

The behaviour of R's core functionality is completely consistent with the idea that the analyst must decide what to do with missing data. A common choice, namely **`leave out records with missing data'** is supported by many base functions through the **na.rm option**.

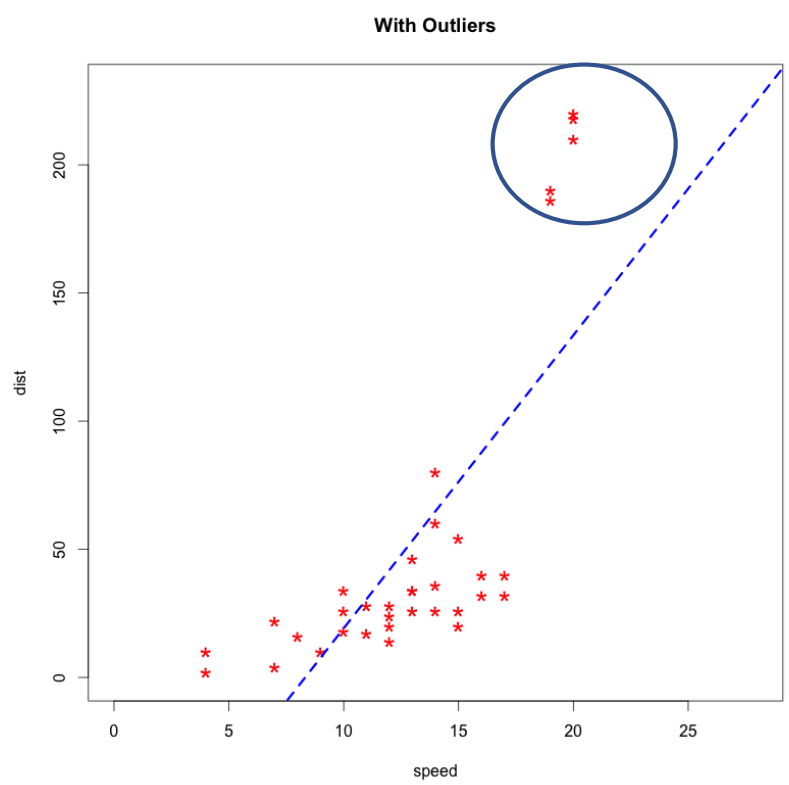
The **complete.cases** function detects rows in a data.frame that do not contain any missing value.

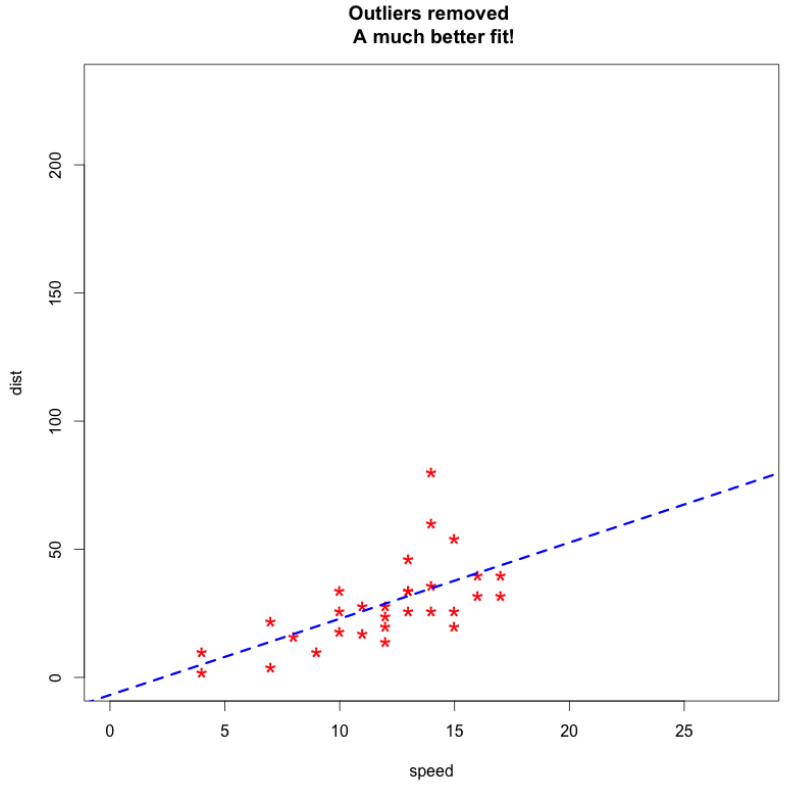
```
age1 <- c(21,42,18, 21 )
height <-c(6,5.9,NA,NA)
data <- data.frame(cbind(age1,height))
complete.cases(data)
(persons_complete <- na.omit(data))
#imputasi mean variabel height
data$height[is.na(data$height)] <- mean(data$height, na.rm = T)
```

```
age1 height
[1,] 21 6.0
[2,] 42 5.9
attr(,"na.action")
 [1] 3 4 attr(,"class")
[1] "omit"
```

Why **outliers** detection is important?

- it can drastically bias/change the fit estimates and predictions.





Handling Outlier univariate using IQR
If data below Q1-1.5IQR, then data is outlier
If data above Q3+1.5IQR, then data is outlier

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
    - faulty data collection instruments
    - data entry problems
    - data transmission problems
    - technology limitation
    - inconsistency in naming convention
- Other data problems which requires data cleaning
    - duplicate records
    - incomplete data
    - inconsistent data

How to Handling Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)
- Regression
  - smooth by fitting the data into regression functions

Data Transformation

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation : Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

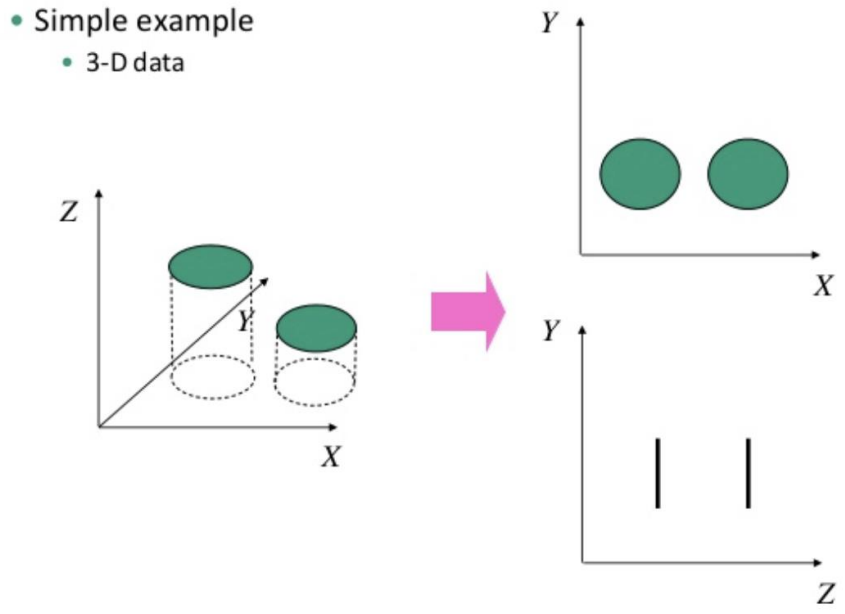
$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1

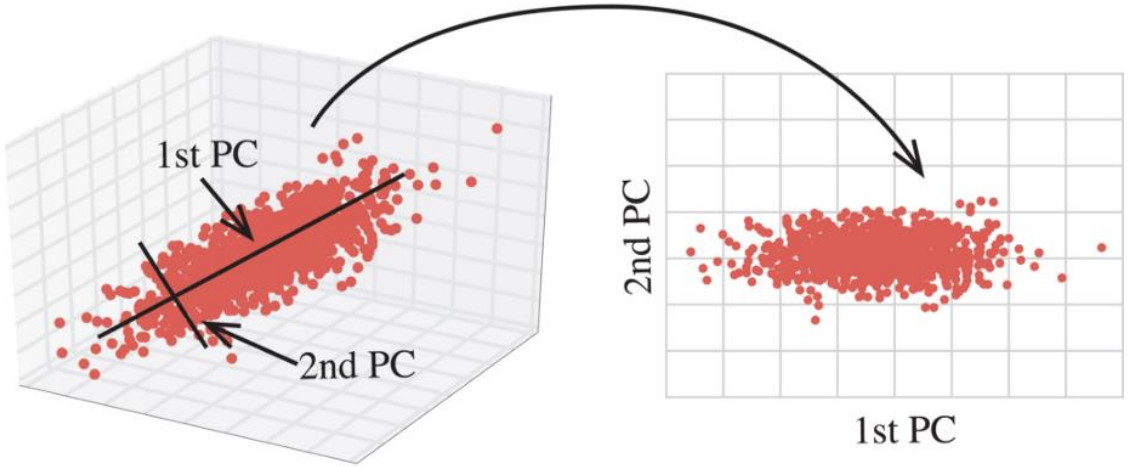

DATA
SCIENCE
INDONESIA

Data Reduction

Data Reduction

- **Dimension reduction** is the process of reducing the number of variables (also sometimes referred to as features or of course dimensions) to a set of values of variables called principal variables.
- Including insignificant variables can significantly impact your **model performance**.





in the first image, it is three dimensional data with **X,Y, Z axes**. The second image is a two dimensional space with **PC1, PC2 as axes.**

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction
  - Recursive Feature Elimination

Dimentionality Reduction

- Feature Extraction
  - Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features).
  - These new reduced set of features should then be able to summarize most of the information contained in the original set of features.
- Method
  - Principal Component Analysis
  - Multidimensional Scaling
  - Independent Component Analysis
  - t-distributed stochastic neighbor embedding (t-SNE)