Detecting outlier and Missing Value- Preprocessing Data

Outline

- Exploratory Data Analysis – Statistics Descriptive
- Data Cleaning
- Data Transformation

Why We Need to Do Preprocessing Data ?

- They are actually from heterogenous sources which have huge size.
- Incomplete Data / Missing Value

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

- Innacurate / Noisy

  containing human error, for example : age = -10

- Inconsistent

  inconsistent in labelling, codes, or names, for example : rating “1,2,3,D,5”

- Duplicate Data

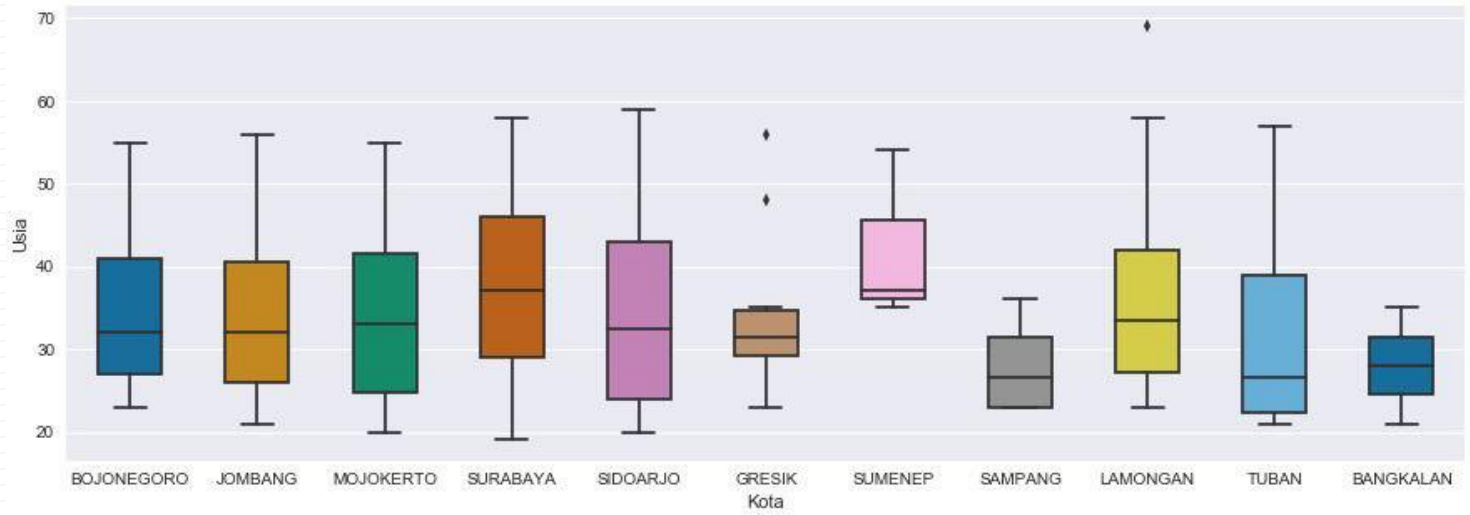
- Cause Incorrect / Misleading Statistics / Bias
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse. —Bill Inmon

Exploratory 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



- different data / outliers and commonly called anomalies where outliers have meaning in a data.
- They should be detected, but not necessarily removed

Data Cleaning

Data Cleaning

- The process of transforming raw data **into consistent data** that can be analyzed.
- **Data cleaning** routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Missing Data

- Data is not available
- 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**

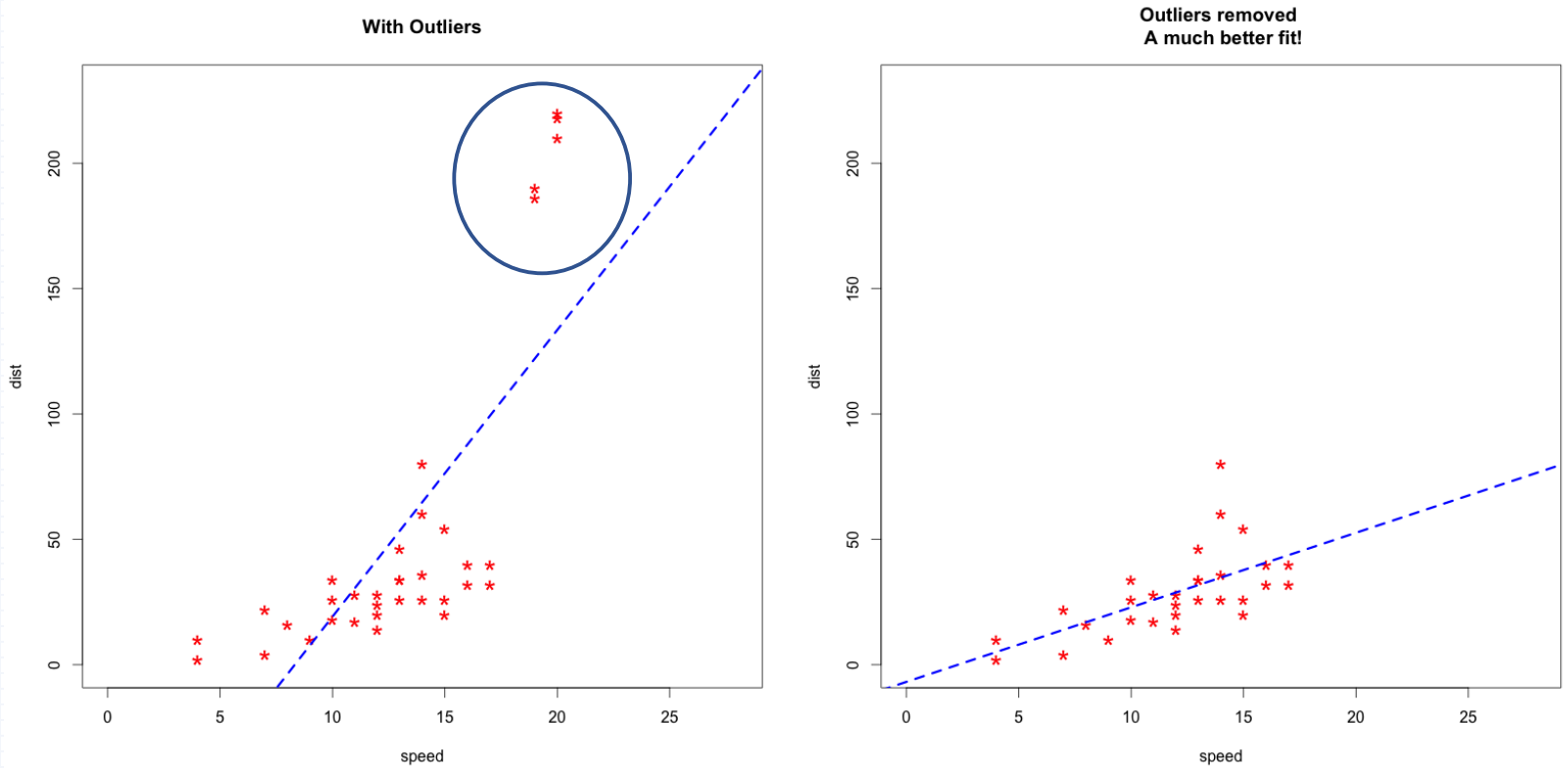
```
>age <- c(23, 16, NA)
> mean(age)
[1] NA
> mean(age, na.rm = TRUE)
[1] 19.5
```

```
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"
```

Outliers detection is important!

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



Data is Outlier Univariate if :

- data below Q1-1.5IQR,
- data above Q3+1.5IQR

IQR (Interquartile Range) = Q3 – Q1

Data Transformation

Data Transformation

- Smoothing : remove noise from data
- Aggregation : Summarization
- Generalization : concept hierarchy climbing
- Normalization : scaled
  - Min-max normalization
  - Z-score normalization
  - Normalization by decimal scaling