Data Cleaning

Week 2

Nurul Afiqah Burhanuddin
nurul.afiqah@ukm.edu.my
Room 2119

Introduction

- Real-world data tend to be dirty, incomplete, and inconsistent.
- Data cleaning is an important step in the data mining process, because quality decisions must be based on quality data.

> “Data of a poor quality are a pollutant of clear thinking and rational decision making. Biased data, and the relationships derived from such data, can have serious consequences in the writing of laws and regulations.”
>
> Hunter (1980)

- Data mining emphasizes data cleansing with respect to the garbage-in-garbage-out principle. Since data mining involves the secondary analysis of large data sets, the dangers are multiplied.
- Data cleaning can improve data quality, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

Characteristics of quality data

1. **Validity.** The degree to which your data conforms to defined business rules or constraints.
2. **Accuracy.** Ensure your data is close to the true values.
3. **Completeness.** The degree to which all required data is known.
4. **Consistency.** Ensure your data is consistent within the same dataset and/or across multiple data sets.

Steps in data cleaning

1. Fix structural errors
2. Handle missing data
3. Handle outliers

Fix structural errors

Some common structural errors across different types of datasets:

- Duplications: one row contains identical information to another row.
- ‘NA’ misclassifications: empty values are misclassified as known values or vice-versa.
- Erroneous observation: incorrect values are entered, either accidentally or deliberately (a common effect of compulsory questions that cannot be answered).
- White space, alphabetical case errors, special character errors, spelling mistakes: values mean the same thing but are classified differently due to white space or alphabetical case.
- Inconsistent time formats: where dates/times have been inputted in several different time formats, making it impossible to make time-based calculations.
- Data recording ambiguity: information is recorded in different ways, e.g., giving a range of values rather than a single value.

Fix structural errors

Some functions from dplyr & tidyr to clean data:

	Function
select a subset of rows	filter()
sort observations based on one variable	arrange()
update or create new columns	mutate()
select a subset of columns	select()
keep only distinct rows	distinct()
separate a character column into multiple columns	separate()
unite multiple columns into one by pasting strings together	unit()

Handling missing data

1. Delete the observation
2. Delete the variable
3. Impute with mean / median / mode
4. Impute by prediction

	Function
kNN	VIM::kNN()
Regression	stats::lm(), stats::predict()
MICE	mice::mice(), mice::complete()

Handling outliers

- An outlier is an observation that differs significantly from other observations.
- Outliers may be due to data entry errors or experimental errors.
- However, if there is no such error, the outlier may indicate something scientifically interesting or rare event.
- Outlier can affect the accuracy of analysis if it is not identified and handled appropriately.

Handling outliers

- How to identify outliers:

1. Visual inspection: scatterplot, boxplot, histogram
2. Statistical tests:

	Function
Grubbs’s test	outliers::grubbs.test()
Rosner’s test	EnvStats::rosnerTest()

3. Modelling:

	Function
Influential data in regression	stats::influence.measures()
Distance from cluster centres	stats::kmeans()

Handling outliers

- After their identification, it is up to your discretion whether to exclude or include them in your analyses.
- It depends on whether the tools/algorithms you will apply are robust to the presence of outliers. For e.g., the slope of a linear model may significantly vary with just one outlier, whereas non-parametric tests are usually robust to outliers.