NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data Cleaning & Preparation

BC2406 UNIT 5

Based on Chew C. H. (2019) textbook: Analytics, Data Science and AI. Vol 1., Chap 5.

Objective

- To clean up the **data quality issues** as best as you can, with the information **currently available**, in order to prepare the data for:
  - More accurate reporting
  - More accurate analysis
  - More accurate models.
- **Impossible** to clean up **perfectly** unless your dataset is small and simple.
  - Iterative. May need to come back to clean further after gaining further information/insights in future.
- May have **more than one way** to clean.
  - **Trade-offs**. Different ways may have different impact in subsequent analysis or models.

Content

- Coding **Missing Values** in Source Data
- Missing Value vs 0 vs Null Value
- Handling (Genuine) Missing Values
- Checking & Correcting **Inconsistencies**

Coding of Missing Values in Source Data

children	Room
1	2
3	3
2	na
0	.
0	3
NA	4
missing	4
N/A	4
m	3
M	2
	4
-99	2
4	3
1	3

There are 9 different ways someone had coded missing values in the two columns.

Use na.strings option to define the 9 human codes for missing value to be NA (R code for missing value).

Note: Rscript to import this dataset and correct the missing value codes provided in ADA1-5-1 lecture.R.

Missing Value vs 0 vs Null Value

- NA: Not Available
  - There exists a value but the value is unknown for now.

- 0: The value is 0 or code for some special value?
  - Affects all the statistics, analysis, reporting and models.
  - Is this a code for something that is not 0 value?
  - Should it really be 0, NA, Null or other value?

- NULL: something that does not exists.
  - R auto-ignore this in many stats, reporting and models.

Handling (Genuine) Missing Values

- Many models (e.g. Linear reg, Logistic reg) auto-ignore rows with NAs.
- Handling NAs
  - Find the correct value and replace the NA.
  - Estimate NA with a value.
  - Use a model that has automatic missing value handling e.g. CART.
  - Delete rows with NA.
    - na.omit() function removes all NAs from the data frame.

To delete or not to delete observations with NAs

- **Delete** as a **last resort**, especially if **a lot of rows has NA**. e.g. is.employed has almost 33% NA.
- What's the reason(s) for NA?
  - Help to determine how to **estimate** the unknown value
  - Missing-at-**random** or missing **systematically**?
- **Type of Variable**: Categorical or Continuous?

Estimating NA in a **Categorical** Variable

- Est. by the **mode** of the column.
- Est. by the mode in the relevant subgroup.
- Est. using a model
  - Logistic Regression
  - CART
  - Others…
- Temporarily recode into another value for consideration in future.

Estimating NA in a **Continuous** Variable

- Missing at Random
  - Est. with mean

- Missing Systematically
  - Est. with mean of the relevant subgroup
  - Est with model
    - Linear Regression
    - CART
    - Others…

- **Discretize** the continuous variable
  - Deal with NAs according to categorical variable.

Verifying the changes after cleaning

- A good habit is to check that changes to data are executed correctly by R.
- Especially if you **overwrite** the data.

Handling Wrong Values

- You know the data value is wrong.
  - Gender = G,
  - Number of Children = 2.1
  - Age = 1098 years
- Handling Wrong Values?
  - Find the **correct** value and replace the wrong value.
  - Estimate wrong value with a **better** value.
  - Replace with **NA** and use a model that has automatic missing value handling e.g. CART.
  - **Delete** rows with wrong values.

Handling Inconsistencies

- When data is recorded **inconsistently**
  - Gender: m/M/F
  - Date format: D-M-Y and M-D-Y
- Find out **cause** of inconsistency
  - Data entry staff did not follow procedure
- If possible, **correct** the cause of the inconsistency instead of data value.
- Some inconsistencies are difficult to correct without further research/questioning.
- Some inconsistencies are harder to detect
  - Domain knowledge
  - Good documented **data dictionary** is very helpful.

Handling Duplicates

- The standard data set assumes each row represent one case or observation.
- However, there could be **duplicate** cases that must be de-duplicated or merged using
  - Identifying information (e.g. NRIC, Invoice Serial Num,…)
  - Definition of a case (aka row) in a dataset depends on context. Are rows 2 and 3 duplicates?

S/N	Name	ID	Date	Outcome
1	Peter Parker	1356	10 Aug 2019	Normal
2	Mary Jane	1455	9 Mar 2019	Normal
3	Mary Jane	1455	11 Aug 2019	Normal

Burden of Proof before taking an action to clean. Is the current data value wrong and can you do <u>better</u>?

1. Some evidence (5%)
2. Reasonable suspicion (10%)
3. Probable cause for arrest (20%)
4. Some credible evidence (30%)
5. Substantial evidence (40%)
6. Preponderance of the evidence (aka balance of probabilities) (51%)
7. Clear and convincing evidence (80%)
8. Beyond reasonable doubt (91%)
9. Beyond a shadow of a doubt (99.9%)
10. Certainty (100%) [Often impossible to achieve]

Summary

- Data Cleaning is **iterative** – can be back to clean later.
- **Logical** Reasoning + **Trade offs**
- NAs in **Categorical** Variables **easier** to treat.
- If using traditional model (e.g. Linear/Logistic Regression), then need to resolve NAs before running regression. Else Regression model will typically delete rows with NA.
- Backup most current dataset before over-writing any columns or any risky data operations. Some prefer to create new columns so as to track changes.
- Always find a way to check that changes was executed correctly as you intended, especially if changes are complex and/or irreversible. i.e. verify.